How human-AI feedback loops alter human perceptual, emotional and social judgements

Moshe Glickman[1, 2, ✉] & Tali Sharot[1, 2, 3, ✉]

[1] Affective Brain Lab, Department of Experimental Psychology, University College London, London, UK

[2] Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London, London, UK

[3] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA

✉ Correspondence authors: mosheglickman345@gmail.com, t.sharot@ucl.ac.uk

Abstract

Artificial intelligence (AI) technologies are rapidly advancing, enhancing human capabilities across various domains spanning from finance to medicine. Despite their numerous advantages, AI systems can exhibit biases in judgments ranging from perception to emotion. Here, in a series of experiments (*N*=1,201), we reveal a feedback loop where human-AI interactions alter processes underlying human perceptual, emotional and social judgements, subsequently amplifying biases in humans. This amplification is significantly greater than observed in interactions between humans, due both to the tendency of AI systems to amplify biases and to how humans perceive AI systems. Participants are often unaware of the extent of the AI's influence, rendering them more susceptible to it. These findings reveal a mechanism wherein AI systems amplify human biases, which are further internalized by humans during human-AI interactions, triggering a snowball effect where small errors in judgment escalate into much larger ones.

Interactions between humans and Artificial Intelligence (AI) technologies have become prevalent, transforming modern society at an unprecedented pace. A vital research challenge is to establish how these interactions alter human beliefs. While decades of research have characterized how humans influence each other (e.g., Centola, 2010; Moussaïd et al., 2017; Zhou et al., 2020), the influence of AI on humans may be qualitatively and quantitively different. This is partially because AI judgments are distinct from human judgements in several ways (for example they tend to be less noisy, Kahneman et al., 2021) and because humans may perceive AI judgements differently from those of other humans (Araujo et al., 2020; Logg et al., 2019). Here, we show how human-AI interactions impact human cognition. In particular, we reveal that when humans repeatedly interact with biased AI systems, they learn to be more biased themselves. We show this in a range of domains and algorithms, including a widely used real-world AI system.

Modern AI systems rely on Machine Learning algorithms to identify complex patterns in vast datasets, without requiring extensive explicit programming. These systems clearly augment human natural capabilities in a variety of domains, such as health care (Hinton, 2018; Loftus et al. 2020; Topol, 2019; Yu et al., 2018), education (Roll et al., 2016), marketing (Ma & Sun, 2020) and finance (Emerson et al., 2019). However, it is well documented that AI systems can automate and perpetuate existing human biases in areas ranging from medical diagnoses to hiring decisions (Caliskan et al., 2017; Obermeyer et al., 2019) and may even amplify those biases (Hall et al., 2022; Leino et al., 2019; Lloyd, 2018). While this problem has been established, a potentially more profound and complex concern has been largely overlooked until now. As critical decisions increasingly involve collaboration between AI and humans (e.g., AI systems assisting physicians in diagnosis, and ChatGPT offering humans advice on various topics, Troyanskaya et al., 2020; Skjuve, 2023), these interactions provide a mechanism through which, not only biased humans generate biased AI systems, but biased AI systems can alter human beliefs, leaving them more biased than they were before. This intuitive possibility, which holds significant implications for our modern society, has not been empirically tested.

Bias, defined as a systematic error in judgments, can emerge in AI systems primarily due to inherent human biases embedded in the datasets the algorithm was trained on ('bias in bias out'; Mayson, 2018, see also Peterson et al., 2022) and/or when the data are more representative of one class than the other (label bias, Buolamwini & Gebru, 2018; Geirhos et al., 2018; Benjamin et al., 2019; Henderson & Serences, 2020). For example, LLM (Large Language Models) systems (e.g., ChatGPT, Bard, Claude) learn from available data on the internet, which being generated by humans contains many inaccuracies and biases, even in cases where the ground truth exists. As a result, these AI systems end up reflecting a host of human biases (such as the conjunction fallacy and the bat-and-ball bias, to name a few, Binz & Schulz, 2023; Yax, Anlló, & Palminteri, 2023). Humans then interact with these LLMs, by asking questions and receiving advice, thus may learn from the models in return. Interaction with other AI systems that exhibit bias (including social bias), such as

text-to-image generative AI models (Luccioni et al., 2023), recommendation algorithms (Morewedge et al., 2023), algorithmic hiring tools (Dastin, 2018), and those that advise humans on credit allocation (Nasiripour, & Natarajan, 2019) and medical care (Ledford, 2019), may also induce similar circularity.

Over a series of studies, we demonstrate that when humans and AI interact, even minute perceptual, emotional and social biases originating either from AI systems or humans leave human beliefs more biased, potentially forming a feedback loop. The impact of AI on humans' beliefs, is gradually observed over time, as humans slowly learn from the AI systems. The amplification effect is greater in human-AI interactions than in human-human interactions, due both to human perception of AI and to the unique characteristics of AI judgements. In particular, AI systems may be more sensitive to minor biases in the data than humans due to their expansive computational resources (Griffiths, 2020) and likely to leverage them to improve prediction accuracy, especially when the data is noisy (Geirhos et al., 2020). Moreover, once trained, AI systems' judgements tend to be less noisy than humans (Kahneman et al., 2021). Thus, AI systems provide a high signal-to-noise ratio than humans, which enables rapid learning by humans, even if the signal is biased. In fact, if the AI is perceived as being superior to humans (as observed in Bogert, Schecter, & Watson, 2021; Hou & Jung, 2021; Logg, Minson, & Moore, 2019, but see Dietvorst, Simmons, & Massey, 2015), learning its bias can be considered perfectly rational. Amplification of bias only occurs if the bias already exists in the system: when humans interact with an accurate AI system their judgements are improved.

Human-AI interactions can create feedback loops that make humans' judgments more biased

We begin by collecting human data in an emotion aggregation task in which human judgement is slightly biased. We then demonstrate that training an AI algorithm on this slightly biased dataset, results in the algorithm not only adopting the bias, but further amplifying it. Next, we show that when humans interact with the biased AI, their initial bias increases (**Fig. 1A,** Human-AI interaction). This bias amplification does not occur in an interaction including only human participants (**Fig. 1B**, Human-Human interaction).

Humans (Level 1) exhibit a small judgment bias. Fifty participants performed an emotion aggregation task (adapted from Haberman et al., 2009; Whitney & Yamanashi Leib, 2017; Goldenberg et al., 2021; Hadar et al., 2022). On each of 100 trials, participants were presented briefly (500ms) with an array of 12 faces and were asked to report whether the mean emotion expressed by the faces in the array was 'more sad' or 'more happy' (**Fig. 1A, Level 1**). The faces were sampled from a data set of 50 morphed faces, created by linearly interpolating between sad and happy expressions (see **Methods**). Based on the morphing ratio, each face was ranked from 1 (100% sad face) to 50 (100% happy face). These ranking were closely associated with participants' own ranking of each face when

observed one-by-one (b = 0.8, $t(50)$ = 26.25, $p < 0.001$, see **Supplementary Results**). We created 100 unique arrays of 12 faces for each participant. The average ranking of the 12 faces in half of the arrays was smaller than 25.5 (thus the array is 'more sad') and greater than 25.5 in the other half (thus the array is 'more happy').

Bias in this task is defined as the difference between the average responses of a participant across all trials and the actual average. The actual average is 0.5, as responses were coded as either 1 ('more sad') or 0 ('more happy'), and exactly half the trials were 'more sad' and half 'more happy'. Mathematically, the bias is expressed as:

$$Bias = \frac{1}{N}\sum_{i=1}^{N} C_i - 0.5$$

Where N denotes the total number of data points, and Ci denotes the classification assigned to each data point (C_i = 1 for a 'more sad' classification and C_i = 0 for a 'more happy' classification). A positive bias indicates a tendency toward classifying responses as 'more sad', while a negative bias, suggests a leaning toward classifying responses as 'more happy'. For example, if a participant classified 0.7 of the arrays as 'more sad' their bias would be 0.7 – 0.5 = 0.2, while a participant who classified 0.3 of the arrays as 'more sad' would have a bias of 0.3 – 0.5 = -0.2.

Consistent with previous studies showing that interpretation of an ambiguous valence is more likely to be negative under short encoding times (Neta & Whalen, 2010; Neta & Tong, 2016), participants showed a slight but significant tendency to report that the faces were 'more sad'. In particular, they categorized 53.08% of the arrays as 'more sad' which is greater than chance (P permutation test against 50% = 0.017, d = 0.34, 95% $CI_{\text{'more sad'}}$ = 0.51-0.56, **Fig. 1E**, green circle; see also **Supplementary Results** for estimation of the bias using a psychometric function analysis). The bias was much larger in the first block than subsequent blocks ($M_{\text{block 1}}$ = 56.72%, $M_{\text{blocks 2-4}}$ = 51.87% P permutation test comparing the first block to the rest = 0.002, d = 0.46, 95% CI = 0.02 to 0.08), suggesting that the participants learn to correct their bias over time.

AI (Level 2) trained on Human judgements from Level 1 amplify human bias. Next, we used a convolutional neural network (CNN; LeCun et al., 2015) to classify each array of faces into 'more happy' or 'more sad'. As detailed below, the CNN amplified the classification bias observed in the human participants (see **Methods** for further details of the model).

First, to test the accuracy of the model, we trained it on the 5,000 arrays that were presented to the participants in Level 1 (5,000 arrays = 50 participants × 100 arrays), with class labels based on the objective ranking scores of the arrays (i.e., *not* the human labels). The model was then evaluated on a 300 out-of-sample test set, and showed classification accuracy of 96%, suggesting it is highly accurate and does not show a bias if trained on non-biased data

(see **Table 1**). Next, we trained the model on class labels defined based on the human classification (5,000 samples of arrays, **Fig. 1A**), and evaluated it on 300 arrays in an out-of-sample test set. The model classified the average emotion as 'more sad' in 65.33% of the cases, despite only 50% of the arrays being 'more sad'. This number was significantly greater than chance (*P* permutation test against 50% < .001, $d = 2.11$, 95% $CI_{\text{'more sad'}} = 0.62\text{-}0.68$, see blue/grey circle in **Fig. 1E**) and also greater than the bias observed in the human data (Level 1) which was only 53% (*P* permutation test < .001, $d = 1.33$, 95% CI = 0.09-0.14, **Fig. 1E**). In other words, the AI algorithm greatly amplified the human bias embedded in the data it was trained on. Similar results were obtained for CNNs with different architectures including ResNet50 (He et al., 2016; see **Supplementary Results**).

A possible reason for the bias amplification of the AI is that it exploits biases in the data to improve its prediction accuracy. This should happen more when data is noisy. To test this hypothesis, we retrained the model with two new sets of labels. First, we used non-noisy labels (i.e., based on the objective ranking scores of the arrays), but induced a minor bias of 3% by switching 3% of the labels. Thus, 53% of the labels were classified as 'more sad'. Second, we used very noisy labels (random labels), in which we also induced a 3% bias. If the bias amplification is due to noise, then the bias of the latter model should be higher than that of the former. The results confirmed this hypothesis (**Table 1**): the average bias of the model trained on the accurate labels with a minor bias was exactly 3%, while the average bias of the model trained on the random labels with a bias of 3% was 50% (i.e., the model classified 100% of arrays as 'more sad'). These results indicate that the bias amplification of the CNN model is related to the noise in the data.

Labels	**Objective ranking** **Acc. = 100%** **Bias = 0%**	**Objective ranking + Minor bias** **Acc. = 97%** **Bias = 3%**	**Participants classifications** **Acc. = 63%** **Bias = 3%**	**Random labels + Minor bias** **Acc. = 50%** **Bias = 3%**
Accuracy – Objective labels	96%	94%	66%	50%
Accuracy – Training labels	96%	92%	69%	53%
Bias	1%	3%	15%	50%

Table 1. Accuracy and bias in the training data and in the CNN's classifications. *Training was conducted using four different label sets: (i) 'Objective' (based on morphing scores), (ii) 'Objective' with a 3% bias (iii) Participants classifications and (iv) Random labels with a 3% bias. The predictions of the model were assessed on an out-of-sample test set of 300 arrays. Accuracy and bias were evaluated with respect to the 'Objective labels' and with respect to the labels the models were trained on ('Training labels'). Acc. = accuracy.*

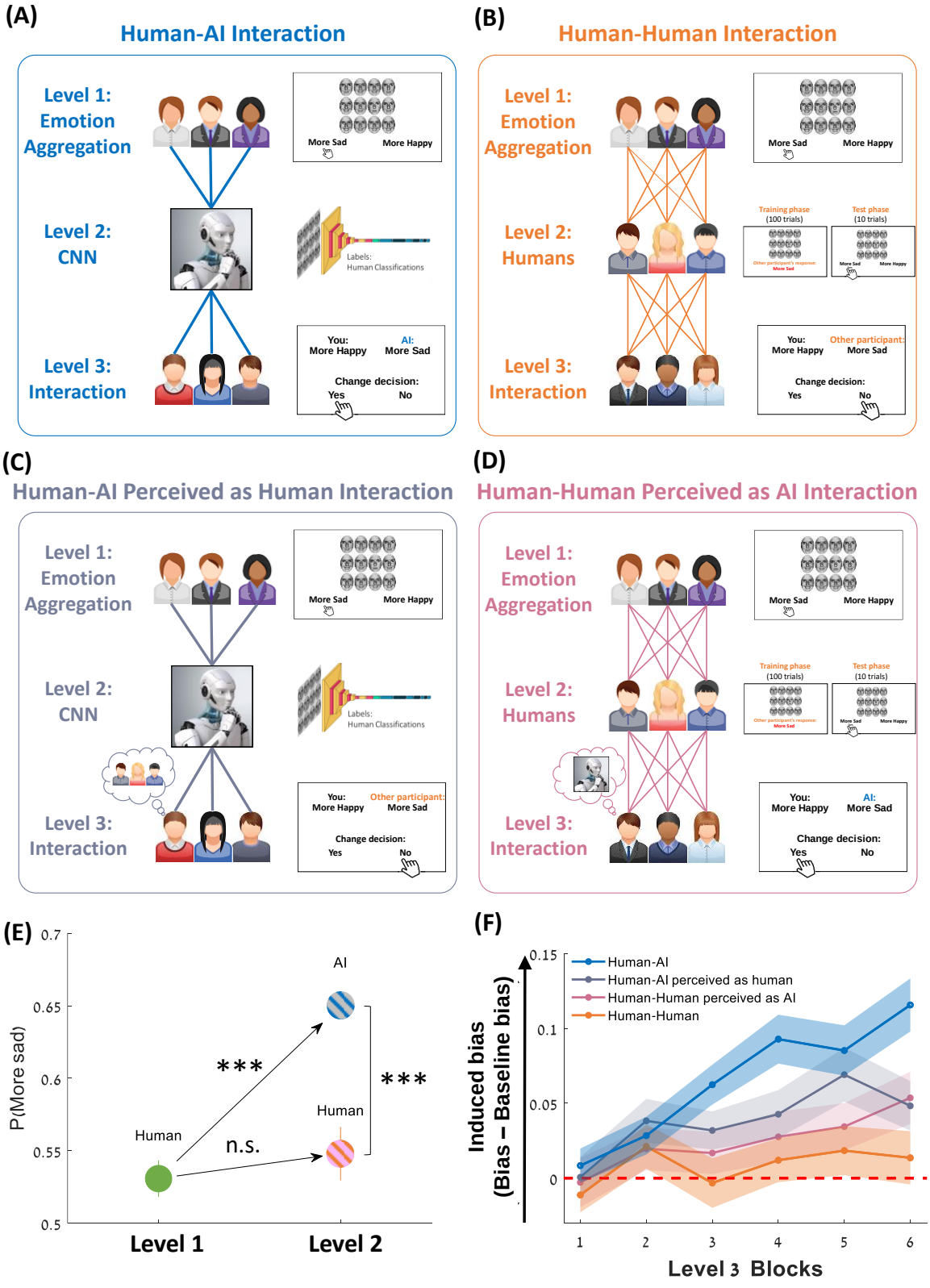

(A)
Human-AI Interaction
Level 1: Emotion Aggregation
Level 2: CNN
Level 3: Interaction
More Sad
More Happy
Labels: Human Classifications
You: More Happy
AI: More Sad
Change decision:
Yes
No
(B)
Human-Human Interaction
Level 1: Emotion Aggregation
Level 2: Humans
Level 3: Interaction
Training phase (100 trials)
Test phase (10 trials)
You: More Happy
Other participant: More Sad
(C)
Human-AI Perceived as Human Interaction
(D)
Human-Human Perceived as AI Interaction
(E)
P(More sad)
AI
Human
***
n.s.
Level 1
Level 2
(F)
Induced bias (Bias – Baseline bias)
Human-AI
Human-AI perceived as human
Human-Human perceived as AI
Human-Human
Level 3 Blocks

Fig. 1. Human-AI interaction creates a feedback loop that make humans more biased (Exp. 1). ***(A) Human-AI interaction****: Human classifications in an emotion aggregation task are collected (Level 1) and fed to an AI algorithm (convolutional neural network; Level 2). A new pool of human participants (Level 3) then interact with the AI.* ***Level 1 (Emotion aggregation):*** *Participants are presented with an array of 12 faces and asked to classify the mean emotion expressed by the faces as 'more sad' or 'more happy'.* ***Level 2 (CNN):*** *The architecture of the CNN used in the experiment. The CNN was trained on human data from Level 1.* ***Level 3 (Human-AI interaction):*** *Participants provide their emotion aggregation response and were then presented with the response of an AI, before being asked if they would like to change their initial response.* ***(B) Human-Human interaction****: Conceptually similar to Human-AI interaction, except that the AI (Level 2) is replaced with human participants. The participants in Level 2 were presented with the arrays and responses of the participants in Level 1 ('Training phase') and then judged new arrays on their own as either 'more sad' or 'more happy' ('Test phase'). The participants in Level 3 were presented with the responses of the human participants from Level 2.* ***(C) Human-AI-perceived-as-human interaction:*** *The condition is also conceptually similar to the Human-AI interaction condition, except that in this condition participants in Level 3 are told they are interacting with another human, while in fact they are interacting with an AI system.* ***(D) Human-Human-perceived-as-AI interaction:*** *The condition is similar to the Human-Human interaction condition, except that participants in Level 3 are told they are interacting with AI, while in fact they are interacting with other humans.* ***(E) Level 1 and 2 results****: Participants in Level 1 (green circle) show a slight bias to respond 'more sad'. This bias is amplified by the AI in Level 2 (blue and grey circle) but not by human participants in Level 2 (orange and pink circle).* ***(F) Level 3 results:*** *When interacting with the biased AI, participants become more biased over time (Human-AI interaction, blue line). In contrast, no bias amplification was observed when interacting with humans (Human-Human interaction, orange line). When interacting with an AI labeled as human (Human-AI-perceived-as-human interaction, grey line) or humans labeled as AI (Human-AI-perceived-as-human interaction, pink line), participants' bias is increased but less than in the Human-AI interaction. Shaded areas correspond to the standard error of the mean. Error bars correspond to standard error of the mean; n.s. = not significant, *** $p < 0.001$.*

Humans (Level 3) interacting with the AI system from Level 2 increase their initial bias. Next, we set out to examine if interacting with the biased AI algorithm would alter human judgments (**Fig. 1A, Level 3**). To this end, we first measured participants baseline performance on the emotion aggregation task for 150 trials, so that we can compare their judgments after interacting with the AI to before. As in Level 1, we found that participants had a small bias at first ($M_{\text{block 1}} = 52.23\%$), which decreased in subsequent blocks, ($M_{\text{blocks 2-5}} = 49.23\%$, P permutation testing first block against the rest of the blocks = 0.03, d = 0.31, 95% CI = 0.01 to 0.06). The question is whether interacting with AI will cause the bias to then reappear in humans and perhaps even increase.

To test this hypothesis, on each of 300 trials, participants first indicated if the array of 12 faces was 'more sad' or 'more happy'. Then they were presented with the response of the AI to the same array (participants were told that they "will be presented with the response

of an AI algorithm that was trained to perform the task"). They were then asked whether they would like to change their initial response or not (that is from 'more sad' to 'more happy' and vice versa). The AI provided different response than the participants on 27.28% (± 1.32% *SE*) of the trials. The participants changed their response on 32.72% (± 2.3% *SE*) of the trials in which the AI provided a different response, and on 0.3% (± 0.1% *SE*) of trials in which the AI provided the same response as they did (these proportions are significantly different: *P* permutation test < 0.001, *d* = 1.97, 95% CI = 0.28 to 0.37). Supplementary study shows that the when not interacting with any associate participants change their decisions only on 3.97% of trials, which is less than when interacting with a disagreeing AI (*P* permutation test < 0.001, *d* =-2.53, 95% CI = -0.57 to -0.42) and more than when interacting with an agreeing AI (*P* permutation test < 0.001, *d* = 0.98, 95% CI = 0.02 to 0.05; see **Supplementary Experiments**).

The primary question of interest, however, is not whether participants changed their response after observing the AI's response. Rather, the critical question is whether over time their *own* response regarding an array (before observing the AI's response to that specific array) became more and more biased due to previous interactions with the AI. That is, did participants learn to become more biased?

Indeed, while in the baseline blocks participants classified on average only 49.9% (± 1.1% *SE*) of the arrays as 'more sad', when interacting with the AI this rate increased significantly to 56.3% (± 1.1% *SE*; *P* permutation test interaction blocks against baseline < 0.001, *d* = 0.84, 95% $CI_{\text{more sad'}}$ = 0.54-0.59). The learned bias increased over time – in the first interaction block it was only 50.72%, whereas in the last interaction block it was 61.44%. The increase in bias across interaction blocks was confirmed by a linear mixed-model predicting 'more sad' classification rate from block number as a fixed factor with random intercepts and slopes at the participant level (b = 0.02, $t(50)$ = 6.23, p < 0.001, **Fig. 1F**).

These results demonstrate an algorithmic bias feedback loop; an AI algorithm trained on a set of slightly biased human data results in the algorithm amplifying it. Subsequent interactions of other humans with this algorithm further increase the humans' initial bias levels, creating a feedback loop.

Bias amplification does not occur in Human-Human interactions

Next, we investigated if the same degree of bias contagion occurs in interactions involving only humans. To this end, we used the same interaction structure as above, except that the AI system was replaced with human participants (**Fig. 1B**).

Humans (Level 1) exhibit small judgement bias. The responses used in the first level of the Human-Human interaction are the same as those used in the Human-AI interaction described above.

Humans (Level 2) trained on Human judgements from Level 1 do not amplify bias. Conceptually similar to AI algorithm training, here we aimed to 'train' humans on human data (**Fig. 1B, Level 2**). The participants were presented with 100 arrays of 12 faces. They were told that they will be presented with the responses of other participants who performed the task before. For each of the 100 arrays, they observed the response of a pseudo-randomly selected participant from Level 1 (see **Methods** for further details). Thereafter they judged 10 new arrays on their own (as either 'more sad' or 'more happy'). To verify that the participants attended to the responses of the other Level 1 participants, they were asked to report it on 20% of the trials (randomly chosen). 14 participants who gave an incorrect answer on more than 10% of the trials (and thus were not attending to the task), were excluded from the experiment.

Participants characterized the arrays as 'more sad' 54.8% of the time, which is different from chance (*P* permutation test against 50% = 0.007, d = 0.41, 95% $CI_{\text{'more sad'}}$ = 52%-58%), but much lower than the AI algorithm which characterized 65.13% of the arrays as 'more sad' (*P* permutation test Level 2 Humans against Level 2 AI < 0.001, d = 0.86, 95% CI = -0.07 to -0.013**, Fig. 1E**).

To examine the generalizability of our findings, we conducted another experiment using a different training method. A new group of participants (N = 50) completed a modified protocol that involved actively predicting the responses of the participants from Level 1. Each correct prediction awarded one point, which was converted to monetary reward at the end of the experiment, thereby incentivizing accurate predictions of other's judgements. The results of this experiment were consistent with those of the previous one: participants characterized the arrays as 'more sad' 53.8% of the time, which is marginally different from chance (*P* permutation test against 50% = 0.08, d = 0.26, 95% $CI_{\text{'more sad'}}$ = 50%-58%), but lower than the AI algorithm (*P* permutation test < 0.001, d = 0.76, 95% CI = -0.07 to -0.015). Thus, the results are robust across different training methods.

To examine if the difference between AI (Human-AI interaction) and humans (Human-Human interaction) in Level 2 is due to humans receiving less training labels, we repeated the same procedure with a new pool of participants (N = 50). However, this time we trained both the humans and the AI system (CNN) on the exact same subset of 200 arrays (see **Methods**). The CNN characterized the arrays as 'more sad' 63.3% of the time, while humans did so only 54.4%. This difference was significant (*P* permutation test < .001, d = 0.69, 95% CI = 0.05 to 0.13). Moreover, the frequency of 'more sad' responses of the human participants who were trained on 200 trials was no different than that of a group of participants who were trained on only 100 trials (M_{100}= 54.8%, M_{200} = 54.4%, *P* permutation test = 0.93, d = 0.03, 95% CI = -0.04 to 0.05), suggesting that human learned bias does not increase with the number of training examples. Together, these results demonstrate that the findings are unlikely to be driven by differences in the training samples sizes.

In conclusion, unlike the AI, human bias was not amplified after being 'trained' on biased human data. This is not surprising, as the level of bias participants in Level 2 naturally exhibit is likely the same as the one they were trained on. Moreover, unlike AI systems, humans base their judgments on factors that go beyond the training session, such as previous experiences and expectations.

Humans (Level 3) interact with Humans from Level 2 do not increase bias. Next, we exposed a new pool of participants (N = 50) to the judgements of humans from Level 2. The task and analysis were identical to that described in Level 3 of the Human-AI interaction (except of course that participants were interacting with humans, which they were made aware of, **Fig. 1B**).

Before being exposed to the other human's response, participants completed five baseline blocks. As in Level 1 and 3 (Human-AI interaction), participants showed a significant bias during the first block ($M_{\text{block 1}}$ = 53.67%) which disappeared over time ($M_{\text{blocks 2-5}}$ = 49.87%, P permutation test first baseline block against the rest of the baseline blocks = 0.007, d = 0.40, 95% CI = 0.01 to 0.06).

Next, participants interacted with other human participants (Human-Human interaction/Level 2). As expected, participants change their classification more when the other participants disagreed with them, 11.27% (± 1.4% *SE*) than when they agreed with them 0.2% (± 0.03% *SE*; P permutation test comparing the two < 0.001, d = 1.11, 95% CI = 0.08 to 0.14) and less than when interacting with a disagreeing AI (which as reminder was 32.72%, P permutation test comparing response change when interacting with a disagreeing AI as compared to interacting with a disagreeing human < 0.001, d = 1.07, 95% CI = 0.16 to 0.27).

Importantly, there was no evidence of learned bias in the Human-Human interaction (**Fig. F**). Classification rates were no different when interacting with other humans ($M_{\text{'more sad'}}$ = 51.45% ± 1.3% *SE*) than baseline (50.6% ± 1.3% *SE*; P permutation test interaction blocks against baseline = 0.48, d = 0.10, 95% $CI_{\text{more sad'}}$ = -0.01 to 0.03) and did not change over time (b = 0.003, $t(50)$ = 1.1, p = 0.27).

Taken together, these results indicate that human bias is significantly amplified in a human-AI interaction, more so than in interaction between humans. The findings suggest that the impact of biased AI systems extends beyond their own biased judgement to their ability to bias human judgment. This poses a concern in any situation where humans interact with an algorithm that may show biases, from interactions with LLMs and text-to-image generative AI systems, to interactions with recommendation algorithms and those designed to support human decisions in different domains such as human resources and medicine.

Is the influence of the AI system driven by the characteristics of its judgement or by human perception of AI?

A question that arises is whether participants became more biased when interacting with the AI system compared to humans because the AI provided more biased judgements, or because they perceive the AI system differently than other humans, or both? To address this question, we ran two additional iterations of the experiment. In the first iteration (AI-perceived-as-human), participants interacted with an AI system but were told they were interacting with another human participant (**Fig. 1C**). In the second iteration (Human-perceived-as-AI), participants interacted with an AI system, but were told they were interacting with another human participant (**Fig. 1D**).

To this end, new pools of participants (N = 50 for each condition) were recruited. First, they performed the baseline test described above and then they interacted with their associate (Level 3). When interacting with the AI (which was believed to be a human) participant's bias increased over time – in the first interaction block it was only 50.5%, whereas in the last interaction block it was 55.28% (**Fig. 1F**). The increase in bias across blocks was confirmed by a linear mixed-model predicting 'more sad' classification rate from block number as a fixed factor with random intercepts and slopes at the participant level (b = 0.01, $t(50)$ = 3.14, $p < 0.001$). Similar results were obtained for the Human-Human-perceived-as-AI interaction. The bias increased across blocks (from 49.0% in the first block to 54.6% in the last), as was confirmed by a linear mixed-model (b = 0.01, $t(50)$ = 2.85, p = 0.004, **Fig. 1F**). In both cases the bias was greater than baseline (Human-AI-perceived-as-Human: M_{bias} = 3.85; P permutation test comparing to baseline = 0.001, d = 0.49, 95% CI = 0.02 to 0.06; Human-Human-perceived-as-AI: M_{bias} = 2.49 ; P permutation test comparing to baseline = 0.04, d = 0.29, 95% CI = 0.01 to 0.05).

Was the induced bias a consequence of the type of input (AI vs. Human) or the perception of that input (perceived as AI vs. perceived as human)? To investigate this, we submitted the induced bias scores (percentage of 'more sad' judgements minus baseline percentage of 'more sad' judgements) into a 2 (Input: AI vs. Human) X 2 (Label: AI vs. Human) Analysis of Variance (ANOVA) with time (blocks 1-6) as a covariate (**Fig. 1F**). The results revealed interactions between input and time: $F(5, 980) = 3.40$, $p = 0.005$ and between label and time: $F(5, 980) = 2.64$, $p = 0.02$. In addition, there was main effect of input: $F(1, 196) = 9.45$, $p = 0.002$, a trend for label: $F(1, 196) = 3.57$, $p = 0.06$, and a main effect of time $F(5, 980) = 14.80$, $p < 0.001$. No other effects were significant (all p's > 0.64). Thus, as illustrated in **Fig. 1F**, both the AI's input and its label contributed to enhanced bias in humans over time.

Finally, we assessed the rate of decision changes among participants. Participants were more likely to change their classification when their associate disagreed with them. In Human-AI-perceived-as-human interactions, decision change occurred at a rate of 16.84% (± 1.2% SE) when there was a disagreement, compared to a mere 0.2% (± 0.05% SE) when agreeing (P permutation test comparing the two < 0.001, d = 1.22, 95% CI = 0.13 to 0.20). Similarly, for the Human-Human-perceived-as-AI condition, decision changes were

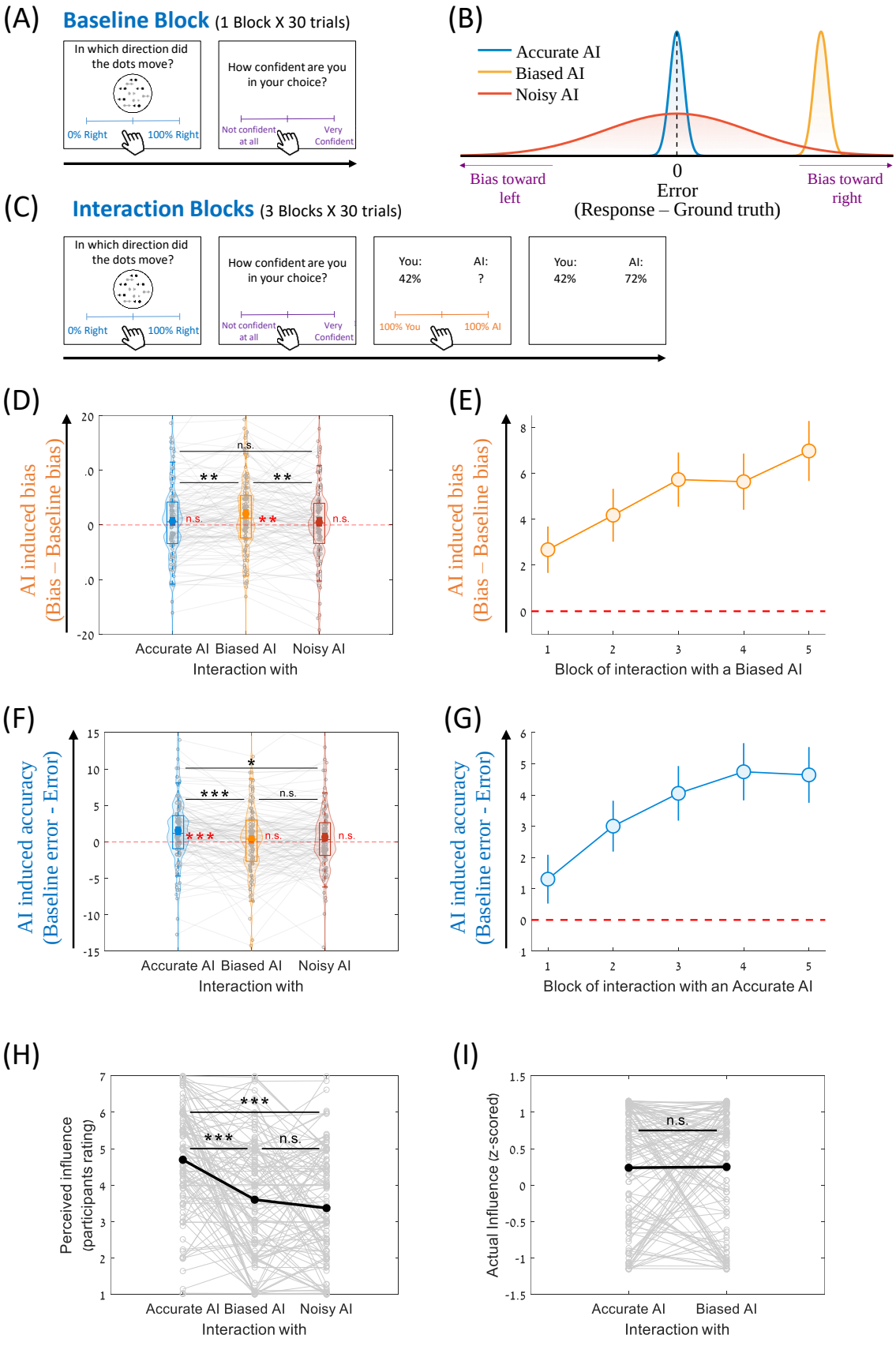
observed in 31.84% (± 2.5% *SE*) when disagreement exists, compared to 0.4% (± 0.1% *SE*) in cases of agreement (*P* permutation test comparing the two < 0.001, *d* = 1.7, 95% CI = 0.26 to 0.36).

To quantify the effects of input and label on decision changes in cases of disagreement, we submitted the percentage of decision change into a 2 (Input: AI vs. Human) X 2 (Label: AI vs. Human) ANOVA with time (blocks 1-6) as a covariate. The results revealed that both the AI's input, $F(1, 196) = 7.05$, $p = 0.009$, and its label, $F(1, 196) = 76.30$, $p < 0.001$, increased the likelihood of decision change. All other main effects and interactions were not significant (all *p*'s > 0.13).

Across different domains, biased algorithms bias human decisions, whereas accurate algorithms improve them.

We next sought to generalize the above results to different types of algorithms and domains. In particular, we aimed to mimic a situation in which humans are not a-priori biased, but rather AI bias emerges for other reasons (for example if it was trained on unbalanced data). To this end, we employed a variant of the Random Dot Kinematogram task (RDK; Bang et al., 2022; Kiani, Hanks, & Shadlen, 2008; Newsome et al., 1989; Liang, Sloane, Donkin, & Newell, 2022), in which participants were presented with an array of moving dots, and were asked to estimate the percentage of dots that move from left to right on a scale ranging from 0% (no dots move from left to right) to 100% (all dots move from left to right). To estimate baseline performance, participants first performed the RDK task on their own for 30 trials and reported their confidence on a scale ranging from 'Not confident at all' to 'Very confident' (**Fig. 2A**). Across trials the actual average percentage of dots that moved rightward was 50.13% ± 20.18 (*SD*), which is not different from 50% (*P* permutation test against 50% = 0.98, *d* = 0.01, 95% CI = 42.93%-57.33%), and the average confidence was 0.56 ± 0.17 (*SD*).

To examine if and how different algorithmic response patterns affect human decision-making, we used three simple algorithms: (i) an accurate algorithm, (ii) a biased algorithm and (iii) a noisy algorithm. The accurate algorithm always indicated the correct percentage of dots that move from left to right (**Fig. 2B** blue distribution). The biased algorithm provided systematically upward biased estimates of dots that move to the right (**Fig. 2B** orange distribution, $M_{bias} = 24.96$). The noisy algorithm provided responses which were equal to those of the accurate algorithm plus Gaussian noise (*SD* = 30; **Fig. 2B** red distribution). The biased and noisy algorithms had the same absolute error (see **Methods** for details). The algorithms used here were hard coded to allow full control over their responses.


(A)
Baseline Block (1 Block X 30 trials)
In which direction did the dots move?
0% Right
100% Right
How confident are you in your choice?
Not confident at all
Very Confident
(B)
Accurate AI
Biased AI
Noisy AI
Bias toward left
0
Error
(Response – Ground truth)
Bias toward right
(C)
Interaction Blocks (3 Blocks X 30 trials)
In which direction did the dots move?
0% Right
100% Right
How confident are you in your choice?
Not confident at all
Very Confident
You: 42%
AI: ?
100% You
100% AI
You: 42%
AI: 72%
(D)
AI induced bias (Bias – Baseline bias)
n.s.
**
**
n.s.
**
n.s.
Accurate AI
Biased AI
Noisy AI
Interaction with
(E)
AI induced bias (Bias – Baseline bias)
Block of interaction with a Biased AI
(F)
AI induced accuracy (Baseline error - Error)
*
***
n.s.
***
n.s.
n.s.
Accurate AI
Biased AI
Noisy AI
Interaction with
(G)
AI induced accuracy (Baseline error - Error)
Block of interaction with an Accurate AI
(H)
Perceived influence (participants rating)
***
***
n.s.
Accurate AI
Biased AI
Noisy AI
Interaction with
(I)
Actual Influence (z-scored)
n.s.
Accurate AI
Biased AI
Interaction with

415

Fig. 2. Biased algorithm produces human bias while accurate algorithm improves human judgment*. (A)* ***Baseline block****. Participants performed the Random Dot Kinematogram (RDK) task, in which an array of moving dots was presented for 1 sec. They estimated the percentage of dots that move from left to right and reported their confidence. (B)* ***Algorithms****. Participants interacted with three algorithms: accurate (blue distribution), biased (orange distribution) and noisy (red distribution). (C)* ***Interaction blocks****. Participant provided their independent judgment and confidence (self-paced) and then observed their own response and a question mark where the AI algorithm response would later appear. Participants were asked to assign weights to their response and the response of the algorithm (self-paced). Thereafter, the response of the algorithm was revealed (2 sec). Note that the AI algorithm's response was revealed only after the participants indicated their weighting. As a result, they had to rely on their global evaluation of the AI based on previous trials. (D)* ***AI induced Bias****. Interacting with a biased AI resulted in significant human bias relative to baseline and relative to interactions with the other algorithms. Circles correspond to the group means; line to median, and the bottom and top edges to the 25th and 75th percentiles. (E) When interacting with a bias algorithm, AI induced biased increases over time. (F)* ***AI induced accuracy change****. Interacting with an accurate AI resulted in significant increase in human accuracy (i.e., reduced error) relative to baseline and relative to interactions with the other algorithms. Notations are the same as in (D). (G) When interacting with an accurate algorithm AI induced accuracy increases over time. (H) Participants perceived the influence of the accurate algorithm on their judgements to be greatest despite (I) that the actual influence of the accurate and biased algorithms was the same. The thin grey lines and circles correspond to individual participants. Error bars correspond to the standard error of the mean * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.*

On each trial participants first provided their judgment and confidence and then observed their own response and a question mark where the algorithm response would later appear (**Fig. 2C**). They were asked to assign weight to their own response and to that of the algorithm on a scale ranging from '100% You' to '100% AI' (see **Methods**). Thus, if a participant assigned a weight of *w* to their own response the final joint decision would be:

$$\textit{Final joint decision} = w\cdot(\textit{participant's response}) + (1\text{-}w)\cdot(\textit{AI's response})$$

This weighting task is analogous to the 'change decision' task in Exp. 1, however, here we use a continuous scale instead of a binary choice, allowing us to get a finer assessment of participants' judgments.

After participants provided their response, the response of the AI algorithm was revealed (**Fig. 2C**). Note that the AI algorithm response was exposed only after the participants indicate their weighting. This was done to prevent participants from relying on the concrete response of the algorithm on a specific trial, and rather rely on their global evaluation of the algorithm. The participants interacted with each algorithm for 30 trials. The order of the algorithms (bias, noisy, accurate) was counterbalanced.

Bias in the RDK task was defined as follows:

$$Bias = \frac{\sum_{i=1}^{n}(Participant's\ response_i - Evidence_i)}{n}$$

where i and n correspond to the index of the present trial and total number of trials, respectively. Evidence corresponds to the percentage of dots that moved rightward in the i-th trial. To compute AI induced bias in participants we subtracted the participant's bias in the baseline block from the bias in the interaction blocks.

$$AI\ induced\ bias = Bias_{AI\ interaction\ blocks} - Bias_{Baseline}$$

At the group level, no systematic bias in baseline responses was detected (Mean response Baseline = 0.62, *P* permutation test against 0 = 0.28, *d* = 0.1, 95% CI = -0.48 to 1.76).

To define accuracy, we first computed an error score for each participant:

$$Error = \frac{\sum_{i=1}^{n}|Participant's\ response_i - Evidence_i|}{n}$$

Then, this quantity was subtracted from the error score in the baseline block, indicating changes in *accuracy*.

$$AI\ induced\ accuracy\ change = Error_{baseline} - Error_{AI\ interaction\ blocks}$$

That is, if errors when interacting with the AI (second quantity) were smaller than baseline errors (first quantity), then the change would be positive, indicating participants became more accurate. However, if errors when interacting with the AI (second quantity) were larger than during baseline (first quantity), then the change would be negative, indicating participants became less accurate when interacting with an AI.

Collaboration was quantified as the average weight assigned to the AI response on a scale ranging from -1 ('100% You', a weight of 0 was assigned to the AI response) to 1 ('100% AI', a weight of 0 was assigned to the AI response).

Results revealed that participants became more biased (towards right) when interacting with the biased algorithm relative to baseline performance ($M\_Bias_{\text{biased AI}}$ = 2.66, M_Bias_{baseline} = 0.62, *P permutation* = 0.002, *d* = 0.28, 95% CI = 0.76-3.35; **Fig. 2D**), and relative to when interacting with the accurate algorithm ($M\_Bias_{\text{accurate AI}}$ = 1.26, *P permutation* = 0.005, *d* = 0.25, 95% CI = 0.42-2.37; **Fig. 2D**) and the noisy algorithm ($M\_Bias_{\text{noisy AI}}$ = 1.15, *P permutation* = 0.006, *d* = 0.25, 95% CI = 0.44-2.56; **Fig. 2D**). No differences in bias were found between the accurate and noisy algorithms, as well as when interacting with these algorithms relative to baseline performance (all *p*'s > 0.29). See also **Supplementary Results** for analysis of the AI induced bias on a trial by trial basis.

The AI induced bias was replicated in a follow-up study (*N* = 50, see **Methods**), in which participants interacted exclusively with a biased algorithm across five blocks (*M_Bias* = 5.03, *P permutation* < 0.001, *d* = 0.72, 95% CI = 3.14 to 6.98, **Fig. 2E**). Critically, we found

a significant linear trend over time (b = 1.0, $t(50)$ = 2.99, p = 0.004, **Fig. 2E**), indicating that the more participants interacted with the biased algorithm, the more biased their judgments became. The learning of the biased induced by the AI was also supported by a computational learning model (see **Supplementary Models**).

Interaction with the accurate algorithm increased the accuracy of participants' independent judgments compared to baseline performance ($M\_Errors_{\text{accurate AI}}$ = 13.48, $M\_Errors_{\text{baseline}}$ = 15.03, $M\_Accuracy\ change_{\text{accurate AI}}$ = 1.55, *P permutation* < 0.001, d = 0.32, 95% CI = 0.69 to 2.42; **Fig. 2F**), and compared to when interacting with the biased algorithm ($M\_Errors_{\text{biased AI}}$ = 14.73, $M\_Accuracy\ change_{\text{biased AI}}$ = 0.03, *P permutation* < 0.001, d = 0.33, 95% CI = 0.58 to 1.94; **Fig. 2F**) and the noisy algorithm ($M\_Errors_{\text{noisy AI}}$ = 14.36, $M\_Accuracy\ change_{\text{noisy AI}}$ = 0.67, *P permutation* = 0.01, d = 0.22, 95% CI = 0.22 to 1.53; **Fig. 2F**). No differences in induced accuracy change were found between the biased and noisy algorithms, as well as no difference in errors when interacting with these algorithms relative to baseline performance (all p's > 0.14, **Fig. 2F**).

The AI induced accuracy change was replicated in a follow-up study (N = 50, see **Methods**), in which participants interacted exclusively with an accurate algorithm across five blocks ($M\_$ *Accuracy change* = 3.55, *P permutation* < .001, d = 0.64, 95% CI = 2.14 to 5.16; **Fig. 2G**). Critically, we found a significant linear trend of the AI induced accuracy change over time (b = 0.84, $t(50)$ = 5.65, p < 0.001, **Fig. 2G**), indicating that the more participants interacted with the accurate algorithm, the more accurate their judgments became.

Participants collaborated more (i.e., assigned a higher weight to the AI response) when interacting with the accurate algorithm, as compared to the biased algorithm ($M\_Collaboration_{\text{accurate AI}}$ = 0.09, $M\_Collaboration_{\text{biased}}$ = -0.09, *P permutation* < 0.001, d = 0.40, 95% CI = 0.10 to 0.26) and the noisy algorithm ($M\_Collaboration_{\text{noisy AI}}$ = 0.09, *P permutation* < 0.001, d = 0.43, 95% CI = 0.10 to 0.25). No differences in collaboration were found between the biased and noisy algorithms (*P permutation* = 0.96, d = 0, 95% CI = -0.05 to 0.05). Consistent with previous results (Liang et al., 2022), our study show that as task difficulty decreased (quantified as the absolute difference between the percentage of dots moving rightward and 50%), participants were less likely to collaborate with the AI algorithms (b = -0.28, $t(120)$ = -5.03, p < 0.001).

Interestingly, participants were more confident when they interacted with the biased algorithm, as compared to the accurate algorithm ($M\_Confidence_{\text{biased AI}}$ = 0.584, $M\_Confidence_{\text{accurate}}$ = 0.558, *P permutation* = 0.003, d = 0.28, 95% CI = 0.01 to 0.04) and noisy algorithm ($M\_Confidence_{\text{noisy AI}}$ = 0.565, *P permutation* = 0.027, d = 0.20, 95% CI = 0.002 to 0.034). No difference in confidence was found between the accurate and noisy algorithms (*P permutation* = 0.37, d = 0.08, 95% CI = -0.01 to 0.03), nor between any of the algorithms and baseline confidence (all p's > 0.18).

Importantly, the increase in accuracy when interacting with the accurate AI could not be attributed to participants 'copying' the algorithm's accurate response. Neither can the increased bias when interacting with the biased algorithm be attributed to participants 'copying' the algorithm's biased responses. This is because we purposefully designed the task such that participants would indicate their judgments on each trial *before* they observed the algorithm's response. Instead, the participants *learned* to provide more accurate judgments in the former case and learned to provide more biased judgements in the latter case.

Participants underestimate the impact of the biased algorithm on their judgment. We were interested in whether participants were aware of the substantial influence the algorithms had on them. To test this, participants were asked to evaluate to what extent they believed their responses were influenced by the different algorithms they interacted with (see **Methods**). As shown in **Fig. 2H**, participants reported being more influenced by the accurate algorithm as compared to the biased (*P permutation* < 0.001, *d* = 0.57, 95% CI = 0.76-1.44) and noisy (*P permutation* < 0.001, *d* = 0.58, 95% CI = 0.98-1.67) algorithms. No significant difference was found between how participants perceived the influence of the biased and noisy algorithms (*P permutation* = 0.11, *d* = 0.15, 95% CI = -0.05 to 0.52).

In reality, however, the magnitude by which they became more biased when interacting with a biased algorithm was equal to the magnitude by which they became more accurate when interacting with an accurate algorithm. We quantified influence using two different methods (see **Methods** section) and both revealed the same result (**Fig. 2I**, Quantifying when Z-scoring across algorithms: *P permutation* = 0.90, *d* = -0.01, 95% CI = -0.19 to 0.17; Quantifying as a percentage difference relative to baseline: *P permutation* = 0.89, *d* = -0.02, 95% CI = -1.44 to 1.9).

These results show that in different paradigms, and under different response protocols, interacting with a biased algorithm biases participants' independent judgments. Moreover, interacting with an accurate algorithm increased the accuracy of participants' independent judgments. Strikingly, the participants were unaware of the strong effect that the biased algorithm had on them.

AI induced bias in humans is generalized to biases in social judgments.

Thus far we demonstrated that interacting with a biased algorithm causes human judgements to be more biased in perceptual and emotion-based tasks. Next, we sought to examine whether the results would generalize to social-based judgments. To investigate this question, we conducted an additional experiment in which participants had to assess the performance of men and women. The question was whether an algorithm that is biased towards men will increase such bias in humans.

Contrary to the moving dots and emotion aggregation tasks, participants may well catch on to the fact that we are assessing bias (in this case gender bias). Thus, they may try to correct their responses to avoid displaying such bias. To overcome this potential problem, we used a two-step task. The first part of the task included direct comparisons between men and women, which make biases relatively apparent. The second part included separate evaluations of men and separate evaluations of women, making biases less apparent (Crandall and Eshleman, 2003). We assumed that in the first part, participants would be aware of the potential gender bias and may suppress it. However, the algorithm's influence from part one may be carried over and emerge in part two, where biases are less apparent and thus less likely to be suppressed.

As in the previous experiment, participants' baseline performance was evaluated prior to interacting with the algorithm ($N = 45$). The first part of the baseline phase (45 trials), was termed 'relative evaluation' (**Fig. 3A**). Participants were told that one of the key skills required for architecture studies is the ability to accurately copy shapes, and that they will be asked to evaluate applicants based on this ability. A simple geometrical shape was presented on screen for one second, next to photos of two applicants and their attempt to copy it. The participants rated which applicant copied the shape more accurately on a scale ranging from one applicant to the other (**Fig. 3A**). Thereafter, participants completed part two – the 'separate evaluation'. In this part, images of all the applicants were presented one-by-one to the participants until response (10 trials). The participant was asked to estimate the likelihood of the applicant to succeed in architecture studies on a scale ranging from 'Not at all' (coded as 1) to 'Very much' (coded as 7).

Participants then proceeded to the interaction phase. First, they performed the same procedure as in the baseline blocks using the same set of photos (**Fig. 3B**). After indicating their response, participants assigned weights to their response and to that of the algorithm on a scale ranging from '100% You' to '100% AI' to determine the final joint decision. Then, they were presented with the algorithm's response (**Fig. 3B**). The responses of the algorithm were hard coded to be accurate when the applicants were of the same gender (man-man or woman-woman trials) but biased towards men otherwise (man-woman trials). We created this algorithm in an attempt to mimic a range of gender-biased algorithm which have been reported in the literature (Dastin, 2018). Specifically, here the algorithm's biased responses were created by overestimating the accuracy of the man applicant in the 'different gender' trials (*Mean bias* = 0.3, *SD* = 0.15). Accuracy was computed based on the correlation between the original and copied shapes (see Methods for details). The copied shapes of men and women did not differ in accuracy ($r_{men} = r_{women} = 0.83$).

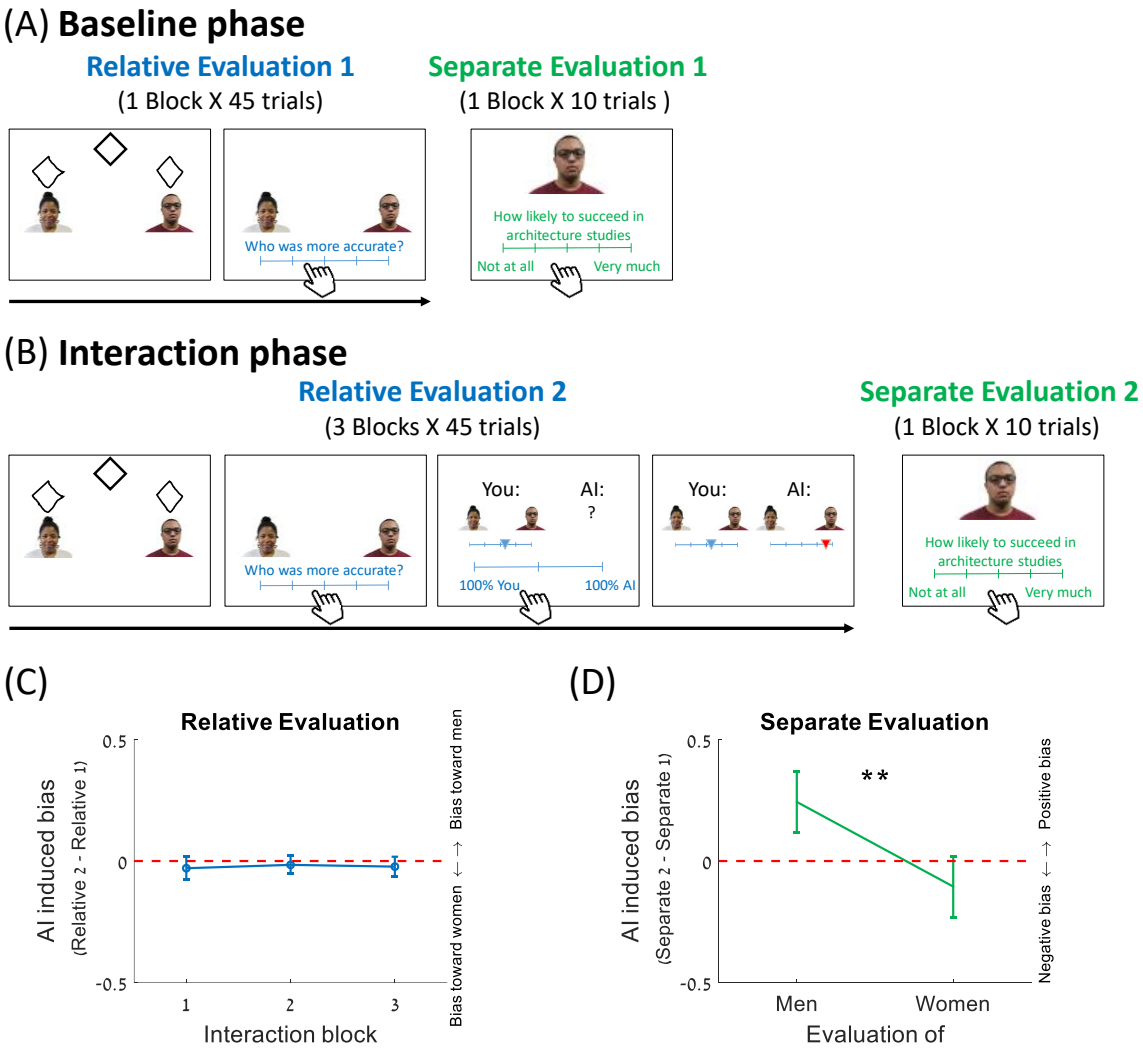


Fig. 3. Gender biased algorithm induces gender bias in humans. *(N = 45). (A)* ***Baseline Phase****. Participants were told that a key skill required for architecture studies is the ability to accurately copy shapes. On each of 45 trials, a simple geometrical shape was presented to them, next to photos of two applicants and their attempt to copy it (1 sec). The participants estimated which applicant copied the shape more accurately (relative evaluation, self-paced). Second, the participants were presented with images of all 10 applicants one-by-one and estimated their likelihood to succeed in architecture studies (separate evaluation, self-paced). (B)* ***AI-Interaction Phase****. Participants first performed the same procedure as in the baseline blocks using the same set of photos. They were then asked to assign weights to their response and the response of the algorithm on a scale ranging from '100% You' to '100% AI' to determine the final joint decision (self-paced). Thereafter, they were presented with the algorithm's response (2 sec), which was accurate for same gender trials but biased towards men otherwise (relative evaluation). Participants then evaluated the applicant one-by-one, exactly as in the baseline phase (separate evaluation, self-paced). (C)* ***No AI induced bias in the relative evaluation task****. Positive values indicate bias toward men after interacting with the AI and negative values indicate bias toward woman. (D)* ***AI induced bias in the separate evaluation task****, such that participants evaluated men as more competent after interacting with the*

*AI relative to before, with no change for evaluation of women. Error bars correspond to standard error of the mean. ** $p < 0.01$.*

We found that participants' separate evaluations of men increased following the collaboration with the biased algorithm (4.42 ± 0.10, *M* ± *SE*) relative to baseline (4.26 ± 0.08, *M* ± *SE*; see **Fig. 3D**), this increase was significantly greater than zero (*P* permutation test = .028, *d* = 0.34, 95% CI = 0.04 to 0.45). The increases in evaluation of men was significantly greater than the non-significant change in the evaluation of women (*P* permutation test = .007, *d* = 0.41, 95% CI = 0.11-0.60). As predicted, the relative evaluations remained steady (*P* permutation test against 0 = .55, *d* = -0.09, 95% CI = -0.10 to 0.04), so that no difference was found after (compared to before) interacting with the algorithm (**Fig. 3C**). These results indicate that interacting with a biased algorithm alters subsequent, independent, evaluations of participants. In particular, participants became more biased in gender-based judgments after interacting with the algorithm, perceiving men applicants as more likely to succeed. The findings also imply that participants may be able to suppress the effect of bias amplification when it was explicitly apparent, but not otherwise.

Amplification of social imbalances by real-world generative AI system biases human judgement.

One approach for providing insight into human cognition is to develop tasks that allow the researcher to manipulate and measure variables of interest in a highly controlled environment (e.g., Botvinick et al., 1999; Daw et al., 2011; Howard & Kahana, 2002). These simplified models of real-world processes provide the experimenter full control over variables, which enables to dissociate their effects. This approach, which we adopted in the above studies, facilitate our ability to measure bias and ground truth more precisely. At the same time, it is valuable to test similar ideas using tools that are in real-world use, despite the inherit loss of experimental control, including less precise measurements.

To that end, in the current experiment we examined changes to human judgements following interactions with Stable Diffusion – a widely used generative AI system designed to create images based on textual prompts (Rombach et al., 2022). Recent studies have reported that Stable Diffusion amplifies existing social imbalances. For example, it over-represents white men in high-power and high-income professions compared to other demographic groups (Bianchi et al., 2023; Luccioni et al., 2023). Such biases can stem from different sources, including problematic training data (Luccioni & Viviano, 2021) and/or flawed content moderation techniques (Luccioni et al., 2023). Stable-Diffusion outputs are used in diverse applications such as videos, advertisements, and business presentations. Consequently, these outputs have the potential to impact humans' belief systems. Here, we test if interacting with Stable Diffusion's outputs increase bias in human judgment.

To test this, we first prompted Stable diffusion to create: "A color photo of a financial manager, headshot, high-quality" (see **Methods**). As expected, the images produced by Stable Diffusion overrepresented white men (85% of images) relative to their representation in the population. For example, in the U.S. only 44.3% of financial managers are men (U.S. Bureau of Labor Statistics, 2022) of which a fraction are white, in the UK only about half are men (Office for National Statistics, 2021) of which a fraction are white. In other Western countries percentage of financial managers who are white men is also less than 85% and in many non-Western countries numbers are likely even lower.

Next, we conducted an experiment (N = 100) to examine how participants' judgments about who is most likely to be a financial manager would alter after interacting with Stable Diffusion. To that end, before and after interacting with Stable Diffusion participants completed 100 trials. On each trial, they were presented with images of six individuals from different race and gender groups: 1) White man, 2) White woman, 3) Asian man, 4) Asian woman, 5) Black man, and 6) Black woman (see **Fig. 4A, Stage 1 - Baseline**). The images were taken from the Chicago Faces Database (Ma et al., 2015), and were balanced in terms of age, attractiveness and racial prototypicality (see **Methods**). On each trial, participants were asked: 'which person is most likely to be a financial manager?'. They responded by clicking on one of the images. Prior to this, participants were provided with a definition of financial manager (see **Methods**). We were interested in whether participants' responses will gravitate towards white men after interacting with Stable Diffusion outputs.

Before interacting with Stable Diffusion, participants selected White men 32.36%, White women 14.94%, Asian men 14.40%, Asian women 20.24%, Black men 6.64% and Black women 11.12% of the time. While there is no definitive ground truth here, based on demographic data, 'White men' is estimated not to be a normative response (for details see **Supplementary Results**). Next, participants were exposed to the outputs of Stable Diffusion (see **Fig. 4A, Stage 2 - Exposure**). Specifically, participants were told that they will be shown three images of financial managers generated by AI (Stable Diffusion) and received a brief explanation about Stable Diffusion (see **Methods**). Then, on each trial, participants viewed three images of financial managers which were randomly chosen from those generated by Stable Diffusion for 1.5 seconds. In stage 3 (**Fig. 4A, Stage 3 – Post-exposure**), participants repeated the task from stage 1. The primary measure of interest was the change in participants' judgements. The data was analyzed using a mixed-model multinomial logistic regression with exposure (before vs. after exposure to AI images) as a fixed factor with random intercepts and slopes at the participant level.

The finding revealed a significant effect for exposure, $F(5, 62) = 5.89$, $p < 0.001$ (**Fig. 4B**), indicating that exposure to the AI images altered human judgements. In particular, exposure increase the likelihood of choosing white men as financial managers ($M_{Before\ exposure}$ = 32.36%, $M_{After\ exposure}$ = 38.20%) compared to: White women ($M_{Before\ exposure}$ =

14.94%, $M_{After\ exposure}$ = 14.40%, b = 0.26, t = 2.08, p = 0.04, 95% CI = 0.01 to 0.50), Asian women ($M_{Before\ exposure}$ = 20.24%, $M_{After\ exposure}$ = 17.14%, b = 0.47, t =3.79, p < 0.001, 95% CI = 0.22 to 0.72), Black men ($M_{Before\ exposure}$ = 6.64%, $M_{After\ exposure}$ = 5.62%, b = 0.65, t = 3.04, p = 0.004, 95% CI = 0.22 to 1.08), Black women ($M_{Before\ exposure}$ = 11.12%, $M_{After\ exposure}$ = 10.08%, b = 0.47, t = 2.46, p = 0.02, 95% CI = 0.09 to 0.87), as well as a trend for Asian men ($M_{Before\ exposure}$ = 14.70%, $M_{After\ exposure}$ = 14.56%, b = 0.28, t = 2.01, p = 0.051, 95% CI = -0.001 to 0.57).

We also run another group of participants to control for order effects. The controls were never exposed to the Stable Diffusion images of financial managers, instead they were exposed to neutral images of fractals (see **Fig. 4A, Stage 2 - Exposure**). The same analysis was performed for the control condition as for the treatment condition. As expected, no significant effect of exposure to neutral fractals was found for the control condition, F(5, 67) = 1.69, p = 0.15 (Fig. 4B). Additionally, no significant differences were observed when comparing White men ($M_{Before\ exposure}$ = 28.42%, $M_{After\ exposure}$ = 27.28%) to each of the demographic groups (all p's > 0.56; White women: $M_{Before\ exposure}$ = 15.64%, $M_{After\ exposure}$ = 15.36%, Asian man: $M_{Before\ exposure}$ = 12.00%, $M_{After\ exposure}$ =11.18%, Asian Women: $M_{Before\ exposure}$ = 20.52%, $M_{After\ exposure}$ = 19.74% and Black men: $M_{Before\ exposure}$ = 8.78%, $M_{After\ exposure}$ = 9.30%), except for an effect in the opposite direction for Black women ($M_{Before\ exposure}$ = 14.64%, $M_{After\ exposure}$ = 17.14%, b = -0.23, t = -2.69, p = 0.06, 95% CI = -0.40 to -0.06). Comparing the treatment and control groups indicates that the former showed a greater increase than the latter in selecting white men after exposure to the images relative to before (P permutation test comparing change in selecting white men across groups = 0.02, d = 0.46, 95% CI = 0.01 to 0.13).

These results suggest that interactions with a commonly used AI system, that amplifies imbalances in real-world representation, induce bias in humans. Crucially, the AI system in this experiment is firmly rooted in the real world. Stable Diffusion has an estimated 10 million users generating millions of images daily (Stability AI, n.d.), underscoring the significance of this phenomenon.

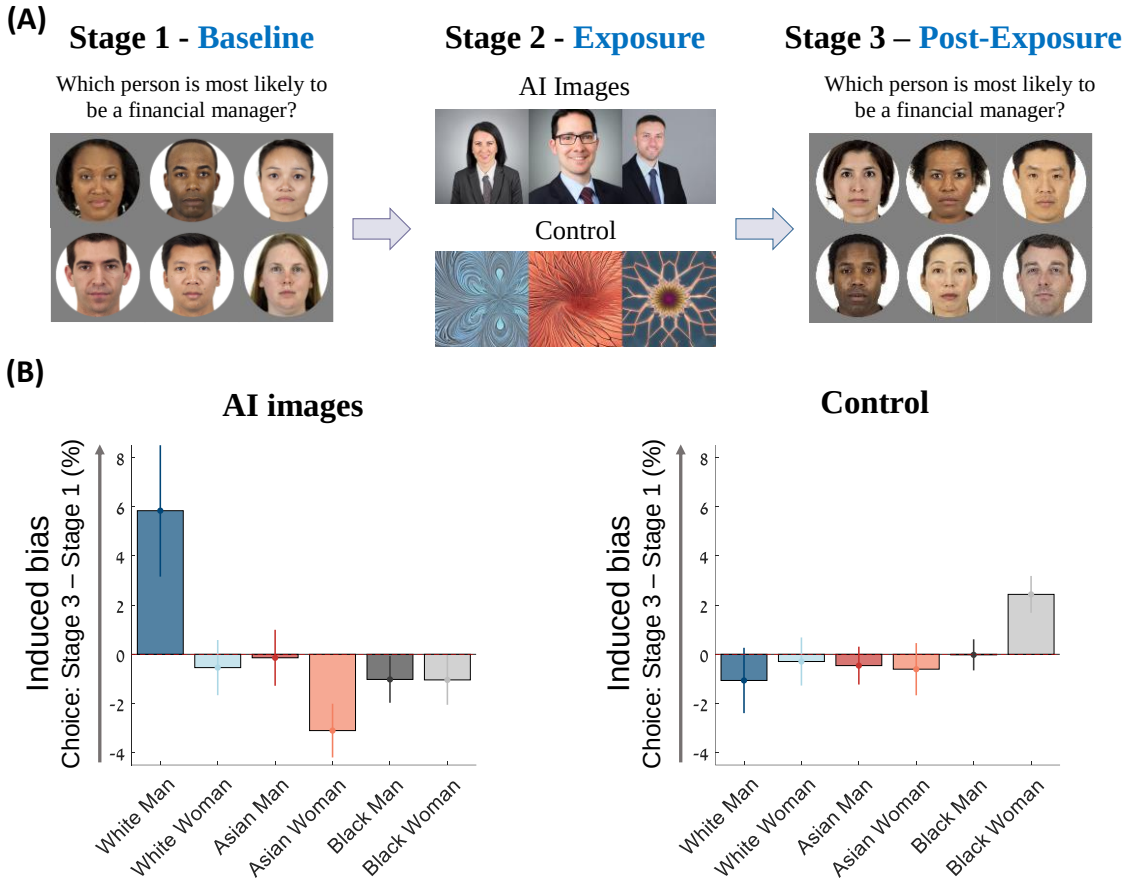


Fig. 4. *Interaction with real-world AI system amplifies human bias (N = 100).* ***(A) Experimental design.*** *The experiment consisted of three stages.* ***Stage 1****: Participants were presented with images featuring six individuals from different race and gender groups: White man, White woman, Asian man, Asian woman, Black man and Black woman. On each trial, participants selected the person who they thought was most likely to be a financial manager.* ***Stage 2****: On each trial, three images of financial managers generated by Stable Diffusion were randomly chosen and presented to the participants. In the control condition, participants were presented with three images of fractals.* ***Stage 3****: Participants repeated the task from stage 1, allowing to measure the change in participants' choices before versus after the exposure to the AI generated images.* ***(B)*** *The results revealed a significant increase in participants' inclination to choose white men as financial managers after being exposed to AI-generated images, but not after being exposed to the fractal neutral images.*

Discussion

Our findings reveal that human-AI interactions create a feedback loop where even small biases emerging from either side increase subsequent human error. First, AI algorithms amplify minute biases embedded in the human data they were trained on. We then reveal that when interacting with biased AI algorithms, humans learned to become more biased. A similar effect was not observed for human-human interactions. Unlike the AI, humans did not amplify the initial small bias present in the data, possibly because humans are less sensitive to minor biases in the data, whereas the AI exploits them to improve its prediction accuracy (see **Table 1**).

The effect of AI induced bias was generalized across a range of tasks and response protocols, including motion discrimination, emotion aggregation and group-based biases. Over time, as participants interacted with the biased AI system repeatedly, their judgments became more biased, suggesting that they learned to adopt the AI systems' bias. Interestingly, participants underestimated the substantial impact of the biased algorithm on their judgment, which could leave them more susceptible to its influence.

We further demonstrated a bias feedback loop in an experiment utilizing a popular real-world AI system – Stable Diffusion. Stable Diffusion tends to over-represent white man when prompted to generate images of high-power/high-income professionals (Luccioni et al., 2023). Here, we show that exposure to such Stable Diffusion images bias human judgement. This likely happens in the real-world when humans prompt Stable Diffusion with similar requests and/or when humans encounter videos or ads generated by Stable Diffusion. Together, this series of experiments unveil a feedback loop that leave humans' beliefs more biased than before interacting with AI systems. As such, the results expose a novel problem that goes beyond important previous findings in AI bias amplification (Zhao et al., 2017; Dinan et al., 2019; Wang & Russakovsky, 2021; Hall et al., 2022; Lloyd, 2018; Leino et al., 2019; Mansoury et al., 2020), AI assisted decision making (Yin, Vaughan, & Wallach, 2019; Lu & Yin, 2021; Cabitza, 2019), impact of AI confidence (Wang et al., 2021; Zhang et al., 2020), Algorithm-in-the-loop (Green & Chen, 2019; De-Arteaga, Fogliato, & Chouldechova, 2020), Human-AI teams (Bansal et al., 2019) and Algorithmic deferral (Keswani, Leae, & Kenthapadi, 2021; Bondi et al., 2022).

The results of our studies underscore the heightened responsibility that algorithm developers must confront in designing and deploying AI systems. Not only may AI algorithms exhibit bias themselves, but they also have the potential to amplify the biases of humans interacting with them, creating a profound feedback loop. The implications can be widespread due to the vast scale and rapidly growing prevalence of AI systems. Of particular concern is the potential effect of biased AIs on children (Kidd & Birhane, 2023),

who have more flexible and malleable knowledge representations, and thus may adopt AI systems' biases more readily. A possibility that has yet to be tested.

It is important to clarify that our findings do not suggest that all AI systems are biased, nor that all AI-human interactions will create a bias. To the contrary, we demonstrate that when humans interact with an accurate AI, their judgments become more accurate (consistent with studies showing that human-AI interaction can improve performance outcomes, e.g., Tschandl et al., 2020). Rather, the results suggest that when a bias exists in the system it has the potential to amplify via a feedback loop. Because biases do exist in both humans and AI systems, this is a problem that should be taken seriously. For example, it has been recently shown that in countries with greater gender inequality, Google search engine is more likely to offer male than female images when searching for the word "person" (Vlasceanu & Amodio, 2022). When this biased collection of images is presented to participants under the (false) claim that those images pop up when searching an unfamiliar profession, such as 'peruker', participants are more likely to indicate they would hire a man as a 'peruker'. It is possible that participants assumed a 'peruker' requires certain traits that are more common in men (such as physical strength or height). Nevertheless, this finding has been offered as evidence for the propagation of human bias by AI. Our results offer clear and direct indication for a human-AI feedback loop that amplifies human bias in domains ranging from perception to emotion detection. Using computational modeling (**Supplementary Models**) we show that humans learn from interactions with an AI algorithm to become biased, rather than just adopting the AI's judgment per-se.

Our results indicate that participants learned the AI system's bias readily, primarily due to the characteristics of the AI's judgments, but also because of participants' perception of the AI (see **Fig. 1F**). Specifically, we observed that when participants were told they were interacting with a human while in fact interacting with an AI, they learn the AI's bias to a lesser extent than when they believed they were interacting with an AI (though they did still significantly learn the bias). This may be because participants perceived the AI systems as superior to humans on the task (as in Bogert, Schecter, & Watson, 2021; Logg, Minson, & Moore, 2019). Thus, participants became more biased, even though they were updating their beliefs in a fashion which may be viewed as perfectly rational. It is also important to note that our results do not suggest that biases will never perpetuate within interactions between humans. But rather, they suggest that when biases are relatively small, they may not do so readily in human-human interactions, as opposed to human-AI interactions. This may be especially true in situations where the evidence is accessible to all (in contrary to 'iterated learning' paradigms for example, Canini et al., 2012).

An intriguing question raised by the current findings is whether the observed amplification of bias endure over time. Further research is required to assess the longevity of this effect. Several factors are likely to influence the persistence of bias, including the duration of

exposure to the biased AI, the salience of the bias and individual differences in the perception of AI systems. For example, a hiring manager who interacts with a biased resume screening system over many selections may show long-lasting bias, whereas a one-time exposure may have more of a transient effect. Nonetheless, even temporary effects could carry significant consequences, particularly considering the scale at which human-AI interactions occur.

In conclusion, AI systems are increasingly integrated into numerous domains, making it crucial to understand how to effectively use them while mitigating their associated risks. The current study reveals that biased algorithms not only produce biased evaluations, but significantly amplify such biases in human judgments, creating a feedback loop. For example, LLMs (such as ChatGPT) or Text-to-Image generators (such as Stable Diffusion) are trained on massive datasets that often contain human introduced biases. As a result, these algorithms tend to reproduce and amplify those biases in their outputs. Interactions with such biased AI systems, whether through queries or image generation, impact users' perceptions and decision-making. This underscores the pressing need to increase awareness of how AI systems influence human judgments. It is possible that strategies aimed at increasing awareness of potential biases induced by AI systems may mitigate their impact, an option that should be tested. Importantly, our results also suggest that interacting with an accurate AI algorithm increases accuracy. Thus, reducing algorithmic bias may hold the potential to reduce biases in humans, increasing the quality of human judgment in domains ranging from health to law.

Methods

Participants. A total of 1,201 individuals participated in this study. Experiment 1 – Level 1: N = 50 (32 women, 18 men, M_{age} = 38.74 ± 11.17 *SD*), experiment 1 – Human-AI – Level 3: N = 50 (24 women, 24 men, 2 not reported, M_{age} = 39.85 ± 14.29 *SD*), experiment 1 – Human-Human – Level 2A: N = 50 (23 women, 25 men, 2 not reported, M_{age} = 34.58 ± 11.87 *SD*), experiment 1 – Human-Human – Level 2B: N = 50 (24 women, 23 men, 1 other, 2 not reported, M_{age} = 36.45 ± 12.97 *SD*), experiment 1 – Human-Human – Level 2C: N = 50 (20 women, 29 men, 1 not reported, M_{age} = 32.05 ± 10.08 *SD*), experiment 1 – Human-Human – Level 3: N = 50 (20 women, 30 men, M_{age} = 40.16 ± 13.45 *SD*), experiment 1 – Human-AI-perceived-as-human – Level 3: N = 50 (15 women, 30 men, 4 not reported, 1 non-binary, M_{age} = 40.16 ± 13.45 *SD*), experiment 1 – Human-Human-perceived-as-AI – Level 3: N = 50 (18 women, 30 men, 1 not reported, 1 non-binary, M_{age} = 34.79 ± 10.80 *SD*), experiment 2: N = 120 (57 women, 60 men, 1 other, 2 not reported, M_{age} = 38.67 ± 13.19 *SD*), experiment 2 accurate algorithm: N = 50 (23 women, 27 men, M_{age} = 36.74 ± 13.45 *SD*), experiment 2 biased algorithm: N = 50 (26 women, 23 men, 1 not reported, M_{age} = 34.91 ± 8.87 *SD*), experiment 3: N = 45 (19 women, 23 men, 1 other, 2 not reported, M_{age} = 39.50 ± 14.55 *SD*), experiment 4: N = 100 (40 women, 56 men, 4

not reported, M_{age} = 30.71 ± 12.07 *SD*), Supplementary experiment 1: N = 50 (26 women, 17 men, 7 not reported, M_{age} = 39.18 ± 14.01 *SD*) and Supplementary experiment 2: N = 386 (241 women, 122 men, 7 other, 16 not reported, M_{age} = 28.07 ± 4.65 *SD*).

Sample sizes were determined based on pilot studies to achieve a power of 0.8 ($\alpha = 0.05$), using g*Power (Faul, Erdfelder, Buchner, & Lang, 2009). In each experiment, the largest N required to detect a key effect was used and rounded up. Participants were recruited via Prolific (https://prolific.ac/) and received in exchange for participation a payment of £7.5 per hour until April 2022, after which the rate was increased to £9 per hour. Additionally, participants in Exps. 1-3 received a bonus fee ranging from £0.5 to £2, which was determined based on performance. All participants had normal or corrected-to-normal vision. All experiments were approved by the UCL Ethics Committee.

Tasks and analyses

1. Emotional Aggregation Task

1.1. AI-Human Interaction

1.1.1. Level 1

Participants performed 100 trials of the emotion aggregation task. On each trial, an array of 12 emotional faces, ranging from sad to happy, was presented for 500ms (**Fig. 1A**). The participants indicated whether, on average, the faces were 'more happy' or 'more sad'. Each participant was presented with 100 unique arrays of faces, which were generated as described in the 'Array of faces' section.

Individual faces. A total of 50 morphed grayscale faces were adopted from Haberman et al., 2009. The faces were created by matching multiple facial features (e.g., corners of the mouth, center of the eye) between extreme sad and happy expressions of the same person (taken from the Ekman gallery, Ekman & Friesen, 1976), and then linearly interpolating between them. The morphed faces ranged from 1 (100% sad face) to 50 (100% happy face), based on the morphing ratio. These objective ranking scores of each face correlated well with participants' subjective perception of the emotion expressed by the face. This was determined by showing participants the faces one-by-one prior to performing the emotion aggregation task, and asking them to rate the faces on a scale from 'very sad' to 'very happy' (self-paced). A linear regression between the 'objective' rankings of the faces and the subjective evaluations of the participants, indicated that the participants were highly sensitive to the emotional expressions (b = 0.8, $t(50)$ = 26.25, $p < 0.001$, R^2 = 0.84).

Array of Faces. The 100 arrays of 12 emotional faces were generated as follows. For 50 of the arrays, the 12 faces were randomly sampled (with repetition) from a uniform distribution in the interval [1,50] with a mean of 25.5. Then, for each of these arrays, a mirror array was created, in which the ranking score of each face was equal to 51 minus

the ranking scores of the face in the original trial. For example, if the ranking scores of faces in an original array were: 21, 44, …, 25, then the ranking scores of the faces in the mirror array were: 51 – 21 = 30, 51 – 44 = 7, …, 51 – 25 = 26. This method ensures that for half the trials the objective mean ranking of the array is higher than the mean of the uniform distribution (mean > 25.5, 'more happy' faces), and in the other half it is lower (mean < 25.5, 'more sad' faces). If the objective mean ranking of an array was exactly 25.5 the faces were resampled.

Bias. Bias in the emotion aggregation task was defined as a percentage of 'more sad' responses beyond 50%. As described in the **Results section**, at the group level the participants showed a tendency to classify the arrays of faces as 'more sad' (*P* permutation test against 50% = 0.017, $d = 0.34$, 95% $CI_{\text{'more sad'}} = 0.51\text{-}0.56$; Neta & Whalen, 2010; Neta & Tong, 2016). Similar results were observed if the bias was quantified using a psychometric function analysis or by using the participants' subjective evaluations of each face instead of their objective ranking (see **Supplementary Results** for more details).

1.1.2. Level 2

The choices of the participants in Level 1 (5,000 choices) were fed into a CNN consisting of five convolutional layers (with filter sizes of 32, 64, 128, 256, 512 and ReLU activation functions) and three fully connected dense layers (**Fig. 1A**). A 0.5 dropout rate was used. The predictions of the CNN were calculated on a test set consisting of 300 new arrays of faces (i.e., arrays that were not included in the training or validation sets). Half of the arrays in the test set had an objective mean ranking score higher than 25.5 (i.e., 'more happy' classification), and the other half had a score lower than 25.5 (i.e., 'more sad' classification).

1.1.3. Level 3

Participants first performed the same procedure described in Level 1, except for performing 150 trials instead of 100. These trials were used to measure baseline performance of participants in the emotion aggregation task. Then, participants performed the emotion aggregation task as in the previous experiment. However, on each trial, after indicating their choice, they were also presented with the response of an AI algorithm for 2sec (**Fig. 1A**). The participants then were asked whether they would like to change their decision (i.e., from 'more sad' to 'more happy' and vice versa) by clicking on the 'Yes' or 'No' buttons (**Fig. 1A**). Before interacting with the AI, participants were told that they "will be presented with the response of an AI algorithm that was trained to perform the task". Overall, participants performed 300 trials divided into six blocks.

1.2. Human-Human Interaction

1.2.1. Level 1

Responses in the first level of the Human-Human interaction were the same as those in the Human-AI interaction.

1.2.2. Level 2

Participants first performed the same procedure as in Level 1. Next, they were presented with 100 arrays of 12 faces for 500ms, followed by the response of another participant from Level 1 to the same array, which was presented for 2sec (**Fig. 1B**). On each trial, the total number of 'more sad' and 'more happy' classifications of the other participants (up until that trial) was presented at the bottom of the screen. Two trials were pseudo-randomly sampled from each of the 50 participants in Level 1. The first trial was sampled randomly and the second was its matched mirror trial. The responses were sampled such that they preserved the bias and accuracy of the full set (with differences in bias and accuracy not exceeding 1%).

To verify that the participants attended to the task on 20% of the trials (randomly chosen) they were asked to report the response of the other player which was presented to them ('what was the response of the other player?' – 'more sad' or 'more happy'). The data of participants whose accuracy scores were lower than 90% were excluded from further analysis (N = 14 participants) for lack of engagement with the task.

After completing this part of the experiment, participants performed the emotion aggregation task again on their own for another 10 trials.

The human participants in the second level were presented with less data than the AI. Thus, it could be argued that they might have shown higher levels of bias if they had been exposed to more data. To test this the same procedure was repeated with new pools of participants (N = 50). However, instead of presenting them 100 responses of the participants in the first Level, they were presented with 200 responses. The responses were semi-randomly sampled, such that they preserved the bias and accuracy of the full set (with differences in bias and accuracy not exceeding 1%).

1.2.3. Level 3

Participants performed the same procedure as described in Human-AI interaction– Level 3, except that here they interacted with a human associate instead of an AI associate. The responses of the human associate were pseudo-randomly sampled from the Human-Human network – Level 2, such that six responses were pseudo-randomly sampled from each participant (a total of 300 trials). Before interacting with the human associate, participants were told that they "will be presented with the responses of another participant who already performed the task".

1.3. Human-AI-perceived-as-human interaction

1.3.1. Level 1

Responses in the first level were the same as those in the Human-AI and Human-Human interactions.

1.3.2. Level 2

The second level was the same as that in the Human-AI interaction.

1.3.3. Level 3

Participants performed the exact same procedure as in the Human-Human interaction. The only difference was, that while they were led to believe that they "will be presented with the responses of another participant who already performed the task", they were in fact interacting with the AI system trained in Level 2.

1.4. Human-Human-perceived-as-AI interaction

1.4.1. Level 1

Responses in the first level were the same as those in the Human-AI and Human-Human interactions.

1.4.2. Level 2

The second level was the same as that in the Human-Human interaction.

1.4.3. Level 3

Participants performed the exact same procedure as in the Human-AI interaction. The only difference was, that while they were led to believe that they "will be presented with the response of an AI algorithm that was trained to perform the task", they were in fact interacting with the human participants from Level 2.

2. ***Random Dot Kinematogram Task***

2.1. Main experiment

2.1.1. Baseline

Participants performed a version of the random dot kinematogram (RDK; Bang et al., 2022; Kiani, Hanks, & Shadlen, 2008; Newsome et al., 1989; Liang, Sloane, Donkin, & Newell, 2022) across 30 trials. On each trial, participants were presented with an array of 100 white dots moving against a gray background. On each trial the percentage of dots moving from left to right were one of the following: 6%, 16%, 22%, 28%, 30%, 32%, 34%, 36%, 38%, 40%, 42%, 44%, 46%, 48%, 50% (presented twice), 52%, 54%, 56%, 58%, 60%, 62%, 64%, 66%, 68%, 70%, 72%, 78%, 86% and 96%. The display was presented for 1sec and then disappeared. Participants were asked to estimate the percentage of dots that moved from left to right on a scale ranging from '0% Left to Right' to '100% Left to Right', as

well as to indicate their confidence on a scale ranging from 'Not confident at all' to 'Very confident' (**Fig. 2A**, upper panel).

2.1.2. Interaction Blocks.

On each trial, participants first performed the RDK task exactly as described above. Then, they were presented with their response (**Fig. 2C**) and a question mark where the AI algorithm response would later appear. They were asked to assign a weight to each response on a scale ranging between '100% You' to '100% AI' (self-paced**).** The final joint response was calculated according to the following formula:

$$\textit{Final joint response} = w\cdot(\textit{participant's response}) + (1\text{-}w)\cdot(\textit{AI's response})$$

Where *w* is the weight the participants assigns to their own response. For example, if the response of the participant was 53% of the dots move rightward and the response of the AI was 73% of the dots move rightward, and the participants assigned a weight of 40% to their response, than the final joint response will be: 0.4·(53%) + 0.6·(73%) = 65% of the dots move rightward. Note that because the AI response was not revealed until the participants indicate their weighting, participants had to rely on their evaluation of the AI based on past trials and could not rely on the response of the AI on that trial. Thereafter the AI response was revealed and remained on screen for 2sec. Participants completed three blocks each consisting of 30 trials.

The participants interacted with three different algorithms: an accurate algorithm, a biased algorithm, and a noisy algorithm (**Fig. 2B**). The accurate algorithm provided the correct response on all trials. The biased algorithm provided a response that was higher than the correct response by 0% to 49% (mean bias 24.96%). The noisy algorithm provided responses which were similar to that of accurate algorithm, but with the addition of significant amount on Gaussian noise (SD_{noise} = 28.46). The error (i.e., mean absolute difference from the correct response) of the biased and noisy algorithms was the same (24.96 and 25.33, respectively).

The order of the algorithms was randomized between participants, using the Latin square method with the following orders: i) accurate, biased, noisy, ii) biased, noisy, accurate and iii) biased, noisy, accurate. Before interacting with the algorithms, participants were told that they "will be presented with the response of an AI algorithm that was trained to perform the task". Before starting each block, participants were told that they would interact with a new and different algorithm. The algorithms were labeled algorithm A, algorithm B and algorithm C. At the end of the experiment, the participants were asked the following questions: i) To what extent were your responses influenced by the responses of Algorithm A? and ii) How accurate was algorithm A? (the questions were repeated for algorithms B and C). The response to the first question was given on a scale ranging from 'Not at all (coded as 1)' to 'Very much (coded as 7)', and to the second question on a scale

ranging from ‘Not accurate at all (coded as 1)’ to ‘Very accurate (coded as 7)’. To assist participants in distinguishing between the algorithms, each algorithm was consistently represented with the same font color (A - green, B - blue, and C - purple) throughout the whole experiment.

We used three main dependent measures: bias, accuracy (error) and collaboration. Bias was defined as the mean difference between a participant’s responses and the correct percentage of dots that moved from left to right. For each participant, the bias in the baseline block was subtracted from the bias in the interaction blocks. The resulting difference in bias was compared against zero. Positive values indicate that participants reported more rightward movement in the interaction blocks than in baseline, while negative values indicate the opposite. Error was defined as the mean *absolute* difference between a participant’s responses and the correct percentage of dots that moved from left to right. In all analyses, for each participant, the error in the interaction blocks was subtracted from the error in baseline blocks. Thus, positive values of this difference score indicated increased accuracy due to interaction with the AI, while negative values indicated reduced accuracy. The tendency to collaborate was defined as the average weight participants assign to the AI response on a scale ranging from -1 (weight of 0% to the AI response) to 1 (weight of 100% to the AI response).

The influence of the biased and accurate algorithms was quantified using two different methods: relative changes and Z-scoring across algorithms. The relative change in bias was computed by dividing the AI induced bias by baseline bias, and the relative change in accuracy was computed by dividing the AI induced accuracy change by baseline error. A comparison of the relative change in bias and in accuracy yielded no significant difference (P permutation = 0.89, d = -0.02, 95% CI = -1.44 to 1.9). The same result was obtained for Z-scoring across algorithms. In this method, we z-scored the AI induced bias of each participant when collaborating with each algorithm (i.e., for each participant we z-scored across algorithms and not across participants). Therefore, three z-scores were obtained for each participant, indicating the relative effect of the biased, accurate and noisy algorithms. The same procedure was repeated for the AI induced accuracy, resulting in three z-scores indicating the relative influence of the different algorithms on the accuracy of each participant. Then, the z-scores of the bias algorithm (for the AI induced bias) and the z-scores of the accurate algorithm (for the AI induced accuracy change) were compared across participants. No significant difference was found between them (P permutation = 0.90, d = -0.01, 95% CI = -0.19 to 0.17).

2.2. Effects Across Time

To examine the AI induced bias and accuracy effects across time, we conducted two additional experiments. In the first one, participants performed the random dot kinematogram (RDK) task exactly as described above, except for one difference. Instead

of interacting with an accurate, biased and noisy algorithms, participants interacted only with a biased algorithm across five blocks. The second experiment was similar to the first one, except for participants interacted with an accurate algorithm across five blocks.

3. Social Judgement Task

In this experiment we examined social-based judgments, specifically gender-based bias. Contrary to the moving dots and emotional aggregation tasks, participants are likely to be aware that we were assessing gender biases. We thus assumed that they would try to correct their responses to avoid displaying them. Thus, to overcome this problem, we adopted a slightly different procedure than in Experiment 1 and 2.

In particular, we introduced a two-step task as detailed below. The first part included direct comparisons between a man and a woman. We assumed that in this part subjects would be aware of the potential gender bias and thus may not show induced bias in their responses. Nevertheless, they may still be impacted by the AI's gender bias and this influence may be revealed in the second part of the experiment. In the second part of the study the subjects evaluated each man and woman separately. Here, it would be difficult for the subjects to track and correct for their own biases, and at this point we may observe the influence of the AI which was induced in the first part of the task. The experiment was as follows:

3.1. Baseline Block

3.1.1. Part I (relative evaluation)

Participants were told that one of the key skills required for architecture studies is the ability to accurately copy geometrical shapes. On each trial, they were presented with photos of two architecture studies applicants next to an image of the work of the applicant – a copy of a shape, which was presented for 1 sec (**Fig. 3A**). Then, the participants were asked to evaluate which applicant copied the shape more accurately, on a scale ranging from '100% applicant A' to '100% applicant B' (self-paced). The original shapes which were used were a square, a square rotated by 45 degrees, a 4-point star, a 4-point star rotated by 45 degrees, a heart, arrow pointing up, down, right and left. The copied shapes were created for the current experiment by manually copying the original shapes using a computer mouse. The total surface of the original and copied shapes was the same.

We used photos of 10 applicants: five men and five women, which were taken from the American Multiracial Faces Database (Chen et al., 2021). The participants completed 45 trials in which they were presented with all possible combinations of pairs of the applicants.

*3.1.2. Part II (separate evaluation***)**

After completing Part I participants were shown all photos again one at a time. They were asked to estimate how likely each applicant is to succeed in architecture studies (self-paced, **Fig. 3A**) on a scale ranging from 'Not at all' to 'Very much' (Separate evaluation 1).

3.2 Interaction Blocks

3.2.1 Part I (relative evaluation)

On each trial, participants first performed the same procedure as in the baseline blocks (**Fig. 3B**). Then, they were presented again with the photos of the two applicants, as well as with their own rating of who copied the original shape more accurately and a question mark where the AI algorithm response would later appear (**Fig. 3B**). As in experiment 2, participants were asked to assign weights to their own response and to that of the algorithm on a scale ranging from '100% You' to '100% AI'. Thereafter the response of the AI algorithm was revealed and presented for 2sec. Before interacting with the AI, participants were told that they "will be presented with the response of an AI algorithm that was trained to perform the task". The participants performed 135 trials divided into 3 blocks. In each block all the possible combinations of pairs of applicants were presented.

The participants in this task interacted only with a gender-biased algorithm. The algorithm provided the accurate responses (see next paragraph for how accuracy was defined) when the comparison was made between two applicants of the same gender (male-male or female-female). However, it was biased in favor of men when the gender of the applicants was different.

The similarity between the original and copied shapes was computed by calculating the correlation between the images. A correlation of 1 indicated that the two images completely overlap, and the higher the differences between the shapes, the lower the correlation between them was. The correlation was optimized for spatial location. The correlations between the original and the copied shapes were converted to a scale ranging from 0 to 1, using min-max normalization:

$$Normalized\ accuracy_i = \frac{Correlation_i - \min(Correlations)}{\max(Correlations) - \min(Correlations)}$$

The AI responses in case of same gender trials (accurate responses) were determined by the difference in the normalized correlations of the two applicants. For example, if applicant A normalized accuracy score was 0.8 and applicant B normalized accuracy score was 0.4, then the AI response was defined as $100 \cdot (0.8 - 0.4) = 40\%$ in favor of applicant A. The AI responses in case of different gender trials (biased responses) were obtained by adding a constant to the correct response in favor of the male applicant in the range of 0.1 to 0.5. The bias was in the range of 0.1-0.42 ($M = 0.3$, $SD = 0.15$).

3.2.2 Part II (Separate evaluation)

After completing Part I participants were shown all photos again one at a time until response. They were asked to estimate how likely each applicant is to succeed in architecture studies on a scale from 'Not at all' to 'Very much' (**Fig. 3B**). AI-induced

gender bias was defined as the mean difference between the average rating given here of the men and women minus these difference in the baseline blocks. Note that biased results in this measure indicate that the gender biased induced by the AI is carried over to other related judgments and may appear even if the participants are trying to correct for bias.

4. Experiment 4

This experiment aimed to investigate whether exposure to images generated by the popular AI system Stable Diffusion (Rombach et al., 2022), which is known to exemplify social imbalances (Luccioni et al., 2023), increases judgment bias in human. To assess this, participants completed a judgement task before and after viewing Stable Diffusion-generated images. Their performance was compared to a control group in which participants were presented with fractals images.

4.1 Procedure

A total of 100 participants were recruited for the experiment. Participants were randomly assigned to either the AI exposure group (N = 50) or to a control fractal exposure group (N = 50).

The study consisted of 3 stages: Stage 1 (Baseline assessment): Participants completed 100 trials in which they were shown an image featuring six individual headshots and were asked: "Who do you think is more likely to be a financial manager?" (see **Fig. 4A**, Stage 1). Participants made their selection by clicking on the chosen image using the computer mouse. Prior to the experiment, participants were provided with a definition of a financial manager (“a person responsible for the supervision and handling of the financial affairs of an organization”, taken from Collins dictionary).

Stage 2 (Exposure): Participants in the AI condition completed 100 trials where they were presented with Stable Diffusion generated images of financial managers (three images per trial). The three images were randomly chosen and presented for 1.5 seconds. Prior to viewing the images, participants were presented with a brief description of Stable Diffusion. Participants in the control group were shown fractal images instead of financial managers images.

Stage 3 (Post-exposure): Participants completed 100 trials repeating the judgement task from Stage 1.

The order of the trials was randomized for all of the stages across participants.

4.2 Stimuli

The stimuli in each trial consisted of images of six individuals (a White man, a White woman, an Asian man, an Asian woman, a Black man and a Black woman) selected from the Chicago Face Database (Ma et al., 2015). From each demographic category, 10 images

of individuals aged 30-40 years were chosen. The chosen individuals were balanced in their age, attractiveness and racial prototypicality (all p's > 0.16). Each image was presented against a grey background with a circle framing the face (see **Fig. 4A**). The locations of the individuals from each demographic group in the image within each trial were randomly determined.

In the AI exposure condition, Stable Diffusion (version 2.1) was used to generate 100 images of financial managers, using the prompt: "A color photo of a financial manager, headshot, high-quality". Images that contained multiple people, unclear faces or distortions were replaced with other images of the same race and gender. The control condition featured 100 fractal images with same size and resolution as the images of the financial managers. Thirty naïve observers categorized the faces according to race and gender (Cohen's kappa = 0.611). Each image was ultimately classified based on the majority categorization across the 30 participants. Of the Stable Diffusion generated images, 85% were classified as White Men, 11% as White Women, 3% as Non-white men and 1% as Non-white women.

Acknowledgements:

We thank B. Blain, I. Cogliati Dezza, R. Dubey, L. Globig, C. Kelly, R. Koster, V. Vellani, S. Zheng, I. Pinhorn, H. Haj-Ali, L. Tse, N. Nachman, R. Moran, M. Usher, I. Fradkin and D. Rosenbaum for critical reading of the manuscript and helpful comments.

Funding: T.S. is funded by a Wellcome Trust Senior Research Fellowship 214268/Z/18/Z.

Author contributions:

Conceptualization: M.G. & T.S.

Methodology: M.G. & T.S.

Investigation: M.G.

Visualization: M.G. & T.S.

Funding acquisition: T.S.

Project administration: T.S.

Supervision: T.S.

Writing – original draft: M.G. & T.S.

Writing – review & editing: M.G. & T.S.

Competing interests: Authors declare that they have no competing interests.

Data and code availability: Data and code are available at: https://github.com/affective-brain-lab/BiasedHumanAI

References

Akaike, H. (1974). A new look at the statistical model identification. IEEE transactions on automatic control, 19(6), 716-723.

Araujo, T., Helberger, N., Kruikemeier, S., & De Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. AI & society, 35, 611-623.

Bang, D., Moran, R., Daw, N. D., & Fleming, S. M. (2022). Neurocomputational mechanisms of confidence in self and others. Nature communications, 13(1), 1-14.

Benjamin, A. S., Qiu, C., Zhang, L. Q., Kording, K. P., & Stocker, A. A. (2019). Shared visual illusions between humans and artificial neural networks. In 2019 Conference on Cognitive Computational Neuroscience (pp. 585-588).

Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., ... & Caliskan, A. (2023, June). Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (pp. 1493-1504).

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. Proceedings of the National Academy of Sciences, 120(6), e2218523120.

Bogert, E., Schecter, A., & Watson, R. T. (2021). Humans rely more on algorithms than social influence as a task becomes more difficult. *Scientific reports*, *11*(1), 1-9.

Botvinick, M., Nystrom, L. E., Fissell, K., Carter, C. S., & Cohen, J. D. (1999). Conflict monitoring versus selection-for-action in anterior cingulate cortex. Nature, 402(6758), 179-181.

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency (pp. 77-91). PMLR.

Canini, K. R., Griffiths, T. L., Vanpaemel, W., & Kalish, M. L. (2014). Revealing human inductive biases for category learning by simulating cultural transmission. Psychonomic Bulletin & Review, 21, 785-793.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183-186.

Centola, D. (2010). The spread of behavior in an online social network experiment. science, 329(5996), 1194-1197.

Crandall, C. S., & Eshleman, A. (2003). A justification-suppression model of the expression and experience of prejudice. Psychological bulletin, 129(3), 414-446.

Crawford, K. (2017, December). The trouble with bias. In Conference on Neural Information Processing Systems, invited speaker.

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. In Ethics of Data and Analytics (pp. 296-299). Auerbach Publications.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. Neuron, 69(6), 1204-1215.

De Martino, B., Kumaran, D., Seymour, B., & Dolan, R. J. (2006). Frames, biases, and rational decision-making in the human brain. Science, 313(5787), 684-687.

Dinan, E., Fan, A., Williams, A., Urbanek, J., Kiela, D., & Weston, J. (2019). Queens are powerful too: Mitigating gender bias in dialogue generation. arXiv preprint arXiv:1911.03842.

Emerson, S., Kennedy, R., O'Shea, L., & O'Brien, J. (2019, May). Trends and applications of machine learning in quantitative finance. In 8th international conference on economics and finance research (ICEFR 2019).

Ekman, P., & Friesen, W. V. (1976). Measuring facial movement. Environmental psychology and nonverbal behavior, 1(1), 56-75.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. Behavior research methods, 41(4), 1149-1160.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231.

Geirhos, R., Jacobsen, J. H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. Nature Machine Intelligence, 2(11), 665-673.

Goldenberg, A., Weisz, E., Sweeny, T. D., Cikara, M., & Gross, J. J. (2021). The crowd-emotion-amplification effect. Psychological science, 32(3), 437-450.

Griffiths, T. L. (2020). Understanding human intelligence through human limitations. Trends in Cognitive Sciences, 24(11), 873-883.

Haberman, J., Harp, T., & Whitney, D. (2009). Averaging facial expression over time. Journal of vision, 9(11), 1-1.

Hadar, B., Glickman, M., Trope, Y., Liberman, N., & Usher, M. (2022). Abstract thinking facilitates aggregation of information. Journal of Experimental Psychology: General, 151(7), 1733.

Hammond, J. S., Keeney, R. L., & Raiffa, H. (1998). The hidden traps in decision making. Harvard business review, 76(5), 47-58.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Henderson, M., & Serences, J. T. (2021). Biased orientation representations can be explained by experience with nonuniform training set statistics. Journal of vision, 21(8), 10-10.

Hinton, G. (2018). Deep learning—a technology with the potential to transform health care. Jama, 320(11), 1101-1102.

Hou, Y. T. Y., & Jung, M. F. (2021). Who is the expert? Reconciling algorithm aversion and algorithm appreciation in AI-supported decision making. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), 1-25.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. Journal of mathematical psychology, 46(3), 269-299.

Huys, Q. J., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., & Roiser, J. P. (2012). Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees. PLoS computational biology, 8(3), e1002410.

Johnson, D. D. (2004). Overconfidence and war: The havoc and glory of positive illusions. Harvard University Press.

Johnson, D. D., & Tierney, D. (2011). The Rubicon theory of war: How the path to conflict reaches the point of no return. International Security, 36(1), 7-40.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). Noise: A flaw in human judgment. Little, Brown.

Kiani, R., Hanks, T. D., & Shadlen, M. N. (2008). Bounded integration in parietal cortex underlies decisions even when viewing duration is dictated by the environment. Journal of Neuroscience, 28(12), 3017-3029.

Kidd, C., & Birhane, A. (2023). How AI can distort human beliefs. Science, 380(6651), 1222-1223.

Klein, J. G. (2005). Five pitfalls in decisions about diagnosis and prescribing. Bmj, 330(7494), 781-783.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

Loftus, T. J., Tighe, P. J., Filiberto, A. C., Efron, P. A., Brakenridge, S. C., Mohr, A. M., ... & Bihorac, A. (2020). Artificial intelligence and surgical decision-making. JAMA surgery, 155(2), 148-158.

Ledford, H. (2019). Millions of black people affected by racial bias in health-care algorithms. Nature, 574(7780), 608-610.

Liang, G., Sloane, J. F., Donkin, C., & Newell, B. R. (2022). Adapting to the algorithm: how accuracy comparisons promote the use of a decision aid. Cognitive research: principles and implications, 7(1), 14.

Lloyd, K. (2018). Bias amplification in artificial intelligence systems. arXiv preprint arXiv:1809.07842.

Luccioni, A. S., Akiki, C., Mitchell, M., & Jernite, Y. (2023). Stable bias: Analyzing societal representations in diffusion models. arXiv preprint arXiv:2303.11408.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes, 151, 90-103.

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. Behavior research methods, 47, 1122-1135.

Ma, L., & Sun, B. (2020). Machine learning and AI in marketing–Connecting computing power to human insights. International Journal of Research in Marketing, 37(3), 481-504.

Mansoury, M., Abdollahpouri, H., Pechenizkiy, M., Mobasher, B., & Burke, R. (2020, October). Feedback loop and bias amplification in recommender systems. In Proceedings of the 29th ACM international conference on information & knowledge management (pp. 2145-2148).

Mayson SG (2019). 'Bias In, Bias Out'. Yale Law Journal, no. 128.

Morewedge, C. K., Mullainathan, S., Naushan, H. F., Sunstein, C. R., Kleinberg, J., Raghavan, M., & Ludwig, J. O. (2023). Human bias in algorithm design. Nature Human Behaviour, 1-3.

Moussaïd, M., Herzog, S. M., Kämmer, J. E., & Hertwig, R. (2017). Reach and speed of judgment propagation in the laboratory. Proceedings of the National Academy of Sciences, 114(16), 4117-4122.

Nasiripour, S., & Natarajan, S. (2019). Apple Co-founder Says Goldman's Apple Card Algorithm Discriminates. Bloomberg. Retrieved from: https://www.bloomberg.com/news/articles/2019-11-10/apple-co-founder-says-goldman-s-apple-card-algo-discriminates.

Neta, M., & Tong, T. T. (2016). Don't like what you see? Give it time: Longer reaction times associated with increased positive affect. Emotion, 16(5), 730.

Neta, M., & Whalen, P. J. (2010). The primacy of negative interpretations when resolving the valence of ambiguous facial expressions. Psychological science, 21(7), 901-907.

Newsome, W. T., Britten, K. H., & Movshon, J. A. (1989). Neuronal correlates of a perceptual decision. Nature, 341(6237), 52-54.

Niv, Y. (2009). Reinforcement learning in the brain. Journal of Mathematical Psychology, 53(3), 139-154.

Norman, G. R., Monteiro, S. D., Sherbino, J., Ilgen, J. S., Schmidt, H. G., & Mamede, S. (2017). The causes of errors in clinical reasoning: cognitive biases, knowledge deficits, and dual process thinking. Academic Medicine, 92(1), 23-30.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

Office for National Statistics. (2021). Retrieved from https://www.bls.gov/cps/cpsaat11.htmhttps://www.ons.gov.uk/aboutus/transparencyandgovernance/freedomofinformationfoi/womeninfinancedirectorpositions

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. Psychological review, 87(6), 532.

Peterson, J. C., Uddenberg, S., Griffiths, T. L., Todorov, A., & Suchow, J. W. (2022). Deep models of superficial face judgments. Proceedings of the National Academy of Sciences, 119(17), e2115228119.

Roll, I., & Wylie, R. (2016). Evolution and revolution in artificial intelligence in education. International Journal of Artificial Intelligence in Education, 26, 582-599.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10684-10695).

Shefrin, H. (2015). How psychological pitfalls generated the global financial crisis. In The Routledge companion to strategic risk management (pp. 289-315). Routledge.

Skjuve, M. Why people use chatgpt. Available at SSRN 4376834.

Stability AI. (n.d.). https://stability.ai/about

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature medicine, 25(1), 44-56.

Troyanskaya, O., Trajanoski, Z., Carpenter, A., Thrun, S., Razavian, N., & Oliver, N. (2020). Artificial intelligence and cancer. Nature Cancer, 1(2), 149-152.

Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., ... & Kittler, H. (2020). Human–computer collaboration for skin cancer recognition. Nature Medicine, 26(8), 1229-1234.

Ubel, P. A. (2009). Free market madness: why human nature is at odds with economics--and why it matters. Harvard Business Press.

U.S. Bureau of Labor Statistics. (2022). Current Employment Statistics Highlights. Retrieved from https://www.bls.gov/cps/cpsaat11.htm

United States Census Bureau. (2022). Retrieved from https://www.census.gov/quickfacts/fact/table/US/PST045219

Vlasceanu M., Amodio D. M. (2022). Propagation of societal gender inequality by internet search algorithms. Proceedings of the National Academy of the Sciences, 119(29), e2204529119.

Wang, A., & Russakovsky, O. (2021, July). Directional bias amplification. In International Conference on Machine Learning (pp. 10882-10893). PMLR.

Wang, D., Zhang, W., & Lim, B. Y. (2021). Show or suppress? Managing input uncertainty in machine learning model explanations. Artificial Intelligence, 294, 103456.

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble perception. Annual review of psychology, 69(1), 105-129.

Yax, N., Anlló, H., & Palminteri, S. (2023). Studying and improving reasoning in humans and machines. arXiv preprint arXiv:2309.12485.

Yu, K. H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. Nature biomedical engineering, 2(10), 719-731.

Zhang, Y., Liao, Q. V., & Bellamy, R. K. (2020, January). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 295-305).

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. arXiv preprint arXiv:1707.09457.

Zhou, B., Pei, S., Muchnik, L., Meng, X., Xu, X., Sela, A., ... & Stanley, H. E. (2020). Realistic modelling of information spread using peer-to-peer diffusion patterns. Nature Human Behaviour, 4(11), 1198-1207.

Zou, J., & Schiebinger, L. (2018). Design AI so that it's fair. Nature, 559(7714), 324-326.

Supplementary Information

Supplementary results

Participants showed high sensitivity to the emotion expressed by the faces

Before performing the emotion aggregation task, participants in Level 1 were presented with 50 morphed faces (adopted from Haberman et al., 2009). The morphed faces were presented one-by-one and ranged from 1 (100% sad face) to 50 (100% happy face). The participants were asked to rate the faces on a scale ranging from 'Very sad' to 'Very happy' (converted to 1 to 50 scale). To examine the sensitivity of the participants to the emotions expressed by the single faces, we conducted a mixed-effects linear regression, predicting the subjective evaluations of the participants from the 'objective' rankings of the morphed faces (i.e., their ranking from 1 to 50, see Methods/Individual faces) as a fixed factor and random intercepts and slopes at the participant level. The regressions revealed that the participants were highly sensitive to the emotional expressions of the faces (b = 0.8, t(50) = 26.25, $p < 0.001$, 95% CI = 0.75 to 0.85).

Estimation of the 'more sad' classification bias using psychometric function analysis

In the main text, bias in the emotion aggregation task was defined by comparing the percentage of 'more sad' responses to chance (50%). In this section, we show that the results are robust to a different bias measure based on a psychometric function analysis. To this end, we performed mixed-model logistic regressions predicting the classifications of the participants (coded as 0 – 'more happy' and 1 – 'more sad') from the mean objective ranking score of each array of 12 faces as a fixed factor and random intercepts and slopes at the participant level. The regressions were conducted separately for the data of the human participants in Level 1, 2 (Human-Human interaction) and 3 (Human-AI interaction, Human-Human interaction, Human-AI-perceived-as-human interaction and Human-Human-perceived-as-AI interaction), as well as for the predictions of the AI in Level 2 (Human-AI interaction).

For each regression model we extracted the indifference point (i.e., the point at which the probability of classifying an array as 'more sad' or 'more happy' is equal to 50%). The mean objective ranking score of each array (scale of 1-50) was converted to a normalized scale ranging from -1 to 1 using the following formula:

$$Normalized\ Mean\ emotion_i = \frac{25.5 - Mean\ emotion_i}{24.5}$$

Thus, an unbiased agent (human or AI) would have an indifference point of 0, whereas a biased agent would have an indifference point higher than 0 (a bias toward 'more sad' responses) or lower than 0 (a bias toward 'more happy' responses). The confidence

intervals (95%) of the indifference points were calculated using bootstrapping method (1,000 bootstraps).

The psychometric function analysis yielded the same results as the ones reported in the main text. The mean indifference point of the human participants in Level 1 was higher than 0 (M = 0.032, 95% CI = 0.018 to 0.047), indicating a bias toward a 'more sad' responses. The mean indifference point of the AI in Level 2 was also higher than 0, and critically higher than that of the participant in Level 1 (M = 0.11, 95% CI = 0.073 to 0.16), indicating that the AI amplified the human participants bias. Finally, the mean indifference point of the participants in Level 3 of the Human-AI condition (M = 0.075, 95% CI = 0.065 to 0.084), was higher than that of the participants in Level 3 of the Human-AI-perceived-as-human condition (M = 0.054, 95% CI = 0.043 to 0.063), the Human-Human-perceived-as-AI condition (M = 0.020, 95% CI = 0.009 to 0.031), as well as of the indifference point of participants in the Human-Human condition (M = 0.016, 95% CI = 0.007 to 0.024).

Convolutional neural network models

In this section, we show that AI bias amplification takes place across different architectures of the convolutional neural networks, including the commonly used ResNet network (He et al. ,2016). The models were trained on the 5,000 arrays that were presented to the participants in Level 1 (5,000 arrays = 50 participants × 100 arrays), with class labels defined based on the human classifications. The models were evaluated using 10 out-of-sample test sets with statistical properties similar to the one used in the main text.

We begin by examining the model described in the main text. As a reminder, this model consisted of five convolution layers with filter sizes of 32, 64, 128, 256 and 512. ReLU activation was used for the convolution layers. The fully connected block consisted of dense layers with 1024, 512 and 128 neurons, along with a softmax activation layer. A 0.5 dropout rate was used. This model showed a mean bias of 11.6% ± 9.0% *SD*.

We explored three additional model architectures. The first included six convolution layers with 32, 64, 128, 256, 512 and 1024 filters. ReLU activation was used for the convolution layers, and the max pooling elements had a size of 2×2. The fully connected block contained two dense layers with 512, 128, along with a softmax activation layer. A 0.5 dropout rate was used. This model showed a mean bias of 9.7% ± 6.7% *SD*.

The second model mirrored the previous one, but the dropout layers were removed. This model showed a mean bias of 13.4% ± 7.1% *SD*.

The third model was based on the ResNet50 architecture (He et al., 2016) to which we added three fully connected layers. This model showed a mean bias of 9.8% ± 10.7% *SD*.

Analysis of the induced bias and accuracy change while controlling for time

The analysis of the data of Exp. 2 did not account for learning effects within the blocks. Therefore, we conducted an additional control analysis of the data of Exp. 2 while controlling for time. To this end, we used two separate mixed model linear regressions. First, we examine if bias alters over the block by conducting a linear mixed model predicting bias on a trial by-trial-basis (defined the difference between the response of the participants and the actual number of dots moving rightward within each trial) from: (i) the type of agent the participant interacted with: biased, noisy and accurate. To represent this categorical variable in the model, two dummy-coded variables were created, using biased agent as the reference category: one dummy variable compared the accurate agent to the biased agent, while the other dummy variable compared the noisy agent to the biased agent. (ii) the evidence (the actual number of dots moving from left to right) and (iii) trial number within each condition (the sequential trial number within each block, ranging from 1 to 30). All were included as fixed factors with random intercepts and slopes. The regression analysis replicated the results of Exp. 2 reported in the main text, showing that participants were more biased when interacting with the bias agent than when interacting with the accurate agent ($b_{Accurate\ vs.\ Biased}$ = -1.38, t(143) = -3.00, p = 0.003, 95% CI = -2.30 to -0.47) and the noisy agent ($b_{Noisy\ vs.\ Biased}$ = -1.50, t(143) = -2.99, p = 0.003, 95% CI = -2.49 to -0.51). Additionally, a significant effect was found for evidence ($b_{Evidence}$ = -0.33, t(120) = -13.22, p < 0.001, 95% CI = -0.38 to -0.28) and for time (b_{Time}= 0.04, t(120) = 2.42, p = .017, 95% CI = 0.008 to 0.08).

Second, we conducted the same analysis as above, but this time predicting error (defined the absolute difference between the response of the participants and the actual number of dots moving rightward within each trial). In this analysis, the type of agent was dummy coded using the accurate agent as the reference category (i.e., one dummy variable compared the biased agent to the accurate agent, while the other dummy variable compared the noisy agent to the accurate agent). Again, the analysis replicated the results of Exp. 2 showing that when interacting with the accurate agent, participants had lower error rates than when interacting with the biased agent ($b_{Biased\ vs.\ Accurate}$ =1.25, $t(153)$ = 3.75, p < 0.001, 95% CI = 0.59 to 1.90) and the noisy agent ($b_{Noisy\ vs.\ Accurate}$ = 0.87, $t(153)$ = 2.79, p = 0.007, 95% CI = 0.24 to 1.50). In addition, there was a significant effect for evidence ($b_{Evidence}$ = -0.05, $t(120)$ = -3.71, p < 0.001, 95% CI = -0.08 to -0.02) and for time (b_{Time} = -0.04, $t(120)$ = -3.79, p < 0.001, 95% CI = -0.07 to -0.02).

Estimating Likelihood of Being a Financial Manager by Demographic Group

In Exp. 4, we presented participants six images of faces from different races and gender groups and asked: '*Who is most likely to be a financial manager?*'. Contrary to Exps. 1-3, in this task there is no objectively 'correct' answer, not least because race and gender are not necessarily good indicators of how likely a person is to be a financial manager. Given, however, that no further information was provided, participants may rely on race and gender to respond. That is, they may have answered the question: '(based on their race and

gender) who is most likely to be a financial manager?'. The answer to this question depends on various factors, such as which country the financial manager works in, and so on. Thus, even for this question there is no definitive ground truth. Nevertheless, we show below that selecting 'white man' is likely not a normative response.

To address the question '*(based on race and gender) Who is most likely to be a financial manager?*' we estimate the probability of being a financial manager given a person's demographic group. We use U.S. as an example due to available statistics from the U.S. Bureau of Labor Statistics (2022), however the principal conclusion likely holds in many other countries. According to this data, among financial managers 44.3% are men and 55.7% are women. Additionally, 78.5% are white, 9.6% are Asian and 9.3% are Black. These percentages do not sum up to 100% because they do not represent all races. Therefore, we treated these groups as a whole, and approximate that among them about 80.6% are White, 9.85% are Asian and 9.55% are Black. Assuming equivalent distribution of men and women across racial group, we assessed that 35.71% of financial managers are White men, 44.89% are White women, 4.36% are Asian men, 5.49% are Asian women, 4.23% are Black men and 5.32% are Black women. Next, we examined the distribution of race and gender in the U.S. (United States Census Bureau., 2022) which is 75.5% White, 6.3% Asian and 13.06% Black and 50.4% women. Therefore based on these demographic data, the probability of being a financial manager given a White man = (Percentage of financial managers that are white men) / (Percentage of population that are white men) = (0.443·0.806·Total number of financial managers)/(0.755·0.496· U.S. population) = 0.95*Total number of financial managers/U.S. population. Doing similar calculations, the results for the other groups are: White women = 1.18*Total number of financial managers/U.S. population, Asian Men = 1.40*Total number of financial managers/U.S. population, Asian Women = 1.73*Total number of financial managers/U.S. population, Black Men = 0.65*Total number of financial managers/U.S. population and Black Women = 0.81*Total number of financial managers/U.S. population. The ratio "Total number of financial managers/U.S. population" remains constant across all demographic groups. As a result, the comparison focuses on the groups' coefficients. Among them, the coefficients of Asian women (1.73), Asian men (1.40) and White women (1.18) are greater than that of White men (0.95). Thus, based purely on demographic group, White men are unlikely to be a normative answer.

Supplementary models

A computational learning model suggest humans learn to be biased from a biased AI. Our tasks in Exp. 2 were designed such that a modulation in the participants' independent judgments indicated that they learned to become more like the algorithm, rather than just mimicking its response. We next use computational modelling to characterize this learning process. To that end, we fitted several reinforcement learning models (Huys et al., 2012;

Niv, 2009) to the RDK data (Exp. 2). First, we examined the performance of a baseline model, which does not assume learning from the algorithms, defined as:

$$Response_t = b_0 + b_1 \cdot Evidence + \varepsilon_t$$

This model postulates that response is a noisy function of the evidence (i.e., the actual percentage of dots moving from left to right). It includes two free parameters: intercept (b_0) and slope (b_1), which map between the participant's internal estimation of the evidence and the external response scale. We compare this baseline model to the following learning model:

$$Response_t = b_0 + b_1 \cdot Evidence_t + Learned\ bias_t + \varepsilon_t$$

$$Learned\ bias_t = Learned\ bias_{t-1} + \alpha \cdot (Response\ AI_{t-1} - Response\ Participant_{t-1})$$

The learning model assumes that in addition to the evaluation of the evidence, a participant's response is also based on the exposure to past responses of the algorithm. In particular, the participant's response (i.e., estimation of the number of dots that move to the right) is assessed against the algorithm's response: If the response of the algorithm is higher than that of the participant (e.g., the algorithm indicates that 75% of dots move to the right, whereas the participant's response is 52%), the participant will tend to overestimate the percentage of right moving dots in the next trial. If, however, the response of the algorithm is lower than that of the participant (e.g., the algorithm indicates the 53% of dots move to the right, whereas the participant's response is 72%), the participant will underestimate the percentage of right moving dots on the next trial. The model includes three parameters: intercept (b_0) and slope (b_1), the weight assigned to the learned bias (b_2) and learning rate parameter (α). The learned bias at t = 1 was set to 0.

In addition to these models, we tested another variant of the learning model assuming that learning rate is modulated by the absolute magnitude of the reward prediction error (hereafter PH learning model; Pearce and Hall, 1980). The Akaike Information Criterion (AIC; Akaike, 1974) scores of the models are presented in **Fig. S1A** (lower valued indicate a better fit). As shown, both learning models decisively outperforms the baseline model. In addition, the PH learning model showed poorer fit than the simpler learning model. The best fitted parameters of the learning model, were all significantly greater than 0: M_{b0} = 0.17, 95% CI 0.15 to 0.20, M_{b1}= .66, 95% CI 0.62 to 0.70, $M\alpha$ = 0.004, 95% CI 0.003 to 0.005. **Fig. S1B** shows the mean response of the participants across trials (blue line), which is well captured by the learning model (black dots). **Fig. S1C** demonstrates that the learning model (*x*-axis) accurately predicts the bias of the participants (*y*-axis, each point represents a single participant) when interacting with the accurate, biased and noisy algorithms.

The learning model suggests that humans learn to be biased over time. This model captures the tendency to produce biased responses while interacting with the biased algorithm, as compared to the accurate and noisy algorithms.

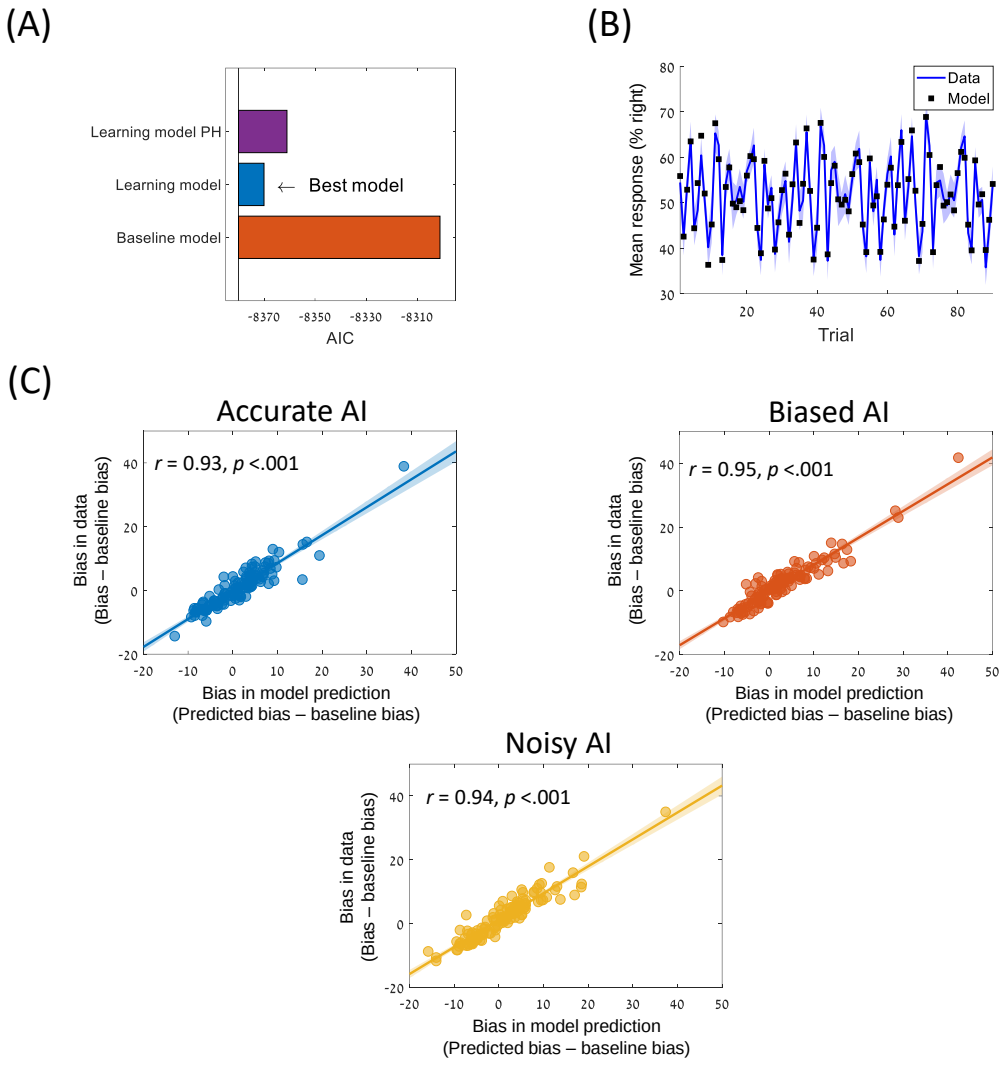


Fig. S1. Computational modeling suggests humans learn to become biased*. (A)* ***Model comparison results****. Both learning model outperformed the baseline model, as indicated by their lower AIC scores. The simpler learning model outperform the PH learning model. (B)* ***Model validation.*** *Predictions of the learning model using the best fit parameters (black dots) shows a good fit to participants mean response (blue line) across trials. Blue shaded area corresponds to 95% confidence interval. (C)* ***Model predicted bias correlates with actual data across conditions.*** *Significant correlations between the bias levels predicted by the learning model (using the best fitted parameters) and the bias of the participants for the accurate (blue), biased (red) and noisy (yellow) algorithms. Each data point represents a single participant.*

Supplementary experiments

Base rate of Response change in Exp. 1

An additional experiment was conducted to investigate the base rate of response change in Exp. 1/Level 3. An overall of 50 participants took part in the experiment (consistent with

the other conditions of Exp. 1). One outlier participant was excluded from analyses (z-score = 6.04, all results remained unchanged even when this participant was included).

The experiment followed a structure similar to that of Exp. 1. Participants first classified arrays of 12 emotional faces as either 'more sad' or 'more happy' in a baseline blocks of 150 trials. Subsequently, they completed 300 trials of the same task, but were given the option to change their initial decisions (by answering 'Yes' or 'No' when prompted). Importantly, unlike the other conditions of Exp. 1, here participants did not interact with any associate (neither AI or human). As in Exp. 1, participants were given a bonus payment based on their final decision, and thus were incentivized to change initial decisions which they considered to be errors.

Analysis of the data showed that the mean decision change rate was 3.97% (± 0.74% *SE*). This rate was significantly lower than in any of the interaction conditions in our study when there were disagreements (all p's < 0.001). Moreover, it was significantly higher than the interaction conditions when there were agreements (all p's < 0.001). The rate of decision change did not vary across blocks, as indicated by a linear mixed model predicting decision changes from block number as a fixed factor with random intercepts and slopes at the participant level ($b = 0.001$, $t(194) = 0.28$, $p = 0.77$).

The low baseline change rate shows that participants rarely change their decisions on their own. Rather, interaction with an associate affects change rates. Specifically, agreement reinforces the initial decision, making changes even less likely, while disagreement increases decision changes compared to baseline rates.

AI induced bias as a function of the bias magnitude

The results of Exp. 2 showed that human participants became more biased when interacting with a biased algorithm. In this section, we examine the association between the magnitude of the bias exhibited by the AI algorithm and the bias induced by it. To this end, we used a paradigm similar to that employed in Exp. 2, with the following exceptions: i) Participants interacted only with a bias algorithm, ii) Participants performed two baseline blocks and three blocks in which they interacted with a biased AI, each of which consisted of 20 trials. The percentage of the dots that moved from left to right were: 6%, 16%, 22%, 24%, 28%, 32%, 36%, 40%, 44%, 48%, 52%, 56%, 60%, 64%, 68%, 72%, 76%, 78%, 86%, 96% (presented in a random order).

Three groups of participants ($N_1 = 127$, $N_2 = 114$, $N_3 = 145$) interacted with three different biased algorithms with average biases of 5.3 (low), 14.5 (medium) and 24.1 (high), respectively (each group interacted only with one algorithm). **Fig. S2A** shows the AI induced bias as a function the magnitude of bias (low, medium and high). One samples *t*-tests against 0, revealed that the AI induced bias was significantly higher than 0 for all

groups: low (M_{bias} = 1.34, $t(126) = 2.38$, $p = 0.019$), medium (M_{bias} = 3.06, $t(113) = 5.34$, $p < 0.001$) and high (M_{bias} = 3.65, $t(144) = 5.95$, $p < 0.001$).

Fig. S2B further shows the AI induced bias of each group as a function of bias magnitude as well as a function of block number. A two-way ANOVA with the bias magnitude (low, medium, high) and Block number (1, 2, 3) as independent variables and AI induced bias as a dependent variable, revealed a main-effect for bias magnitude, $F(2, 383) = 4.23$, $p = 0.015$. Follow-up post-hoc tests, indicate that the AI induced biased was higher when bias magnitude was high compared to low ($p = 0.01$). A trend in the same direction was also found when comparing the medium and low groups ($p = 0.09$). No difference was found between the high and medium groups ($p = 0.484$). A main-effect was found also for block, $F(2, 383) = 9.78$, $p < 0.001$. Follow-up post-hoc tests, showed that the AI induced biased was higher in the third ($p < 0.001$) and second ($p < 0.001$) blocks as compared to the first one. No difference was found between the second and third blocks ($p = 0.309$). A linear trend analysis confirmed that overall, the AI induced bias increased across blocks ($t(383) = 4.23$, $p < 0.001$). The interaction between bias magnitude and block number did not reach a statistical significance, $F(2, 383) = 1.79$, $p = 0.128$.

These results replicate the AI induced bias effect shown in Exp. 2. The results also replicate the increase of the AI induced bias across blocks. Moreover, the results show that even if the bias of the AI is as low as 5%, the participants are still significantly influenced by it.

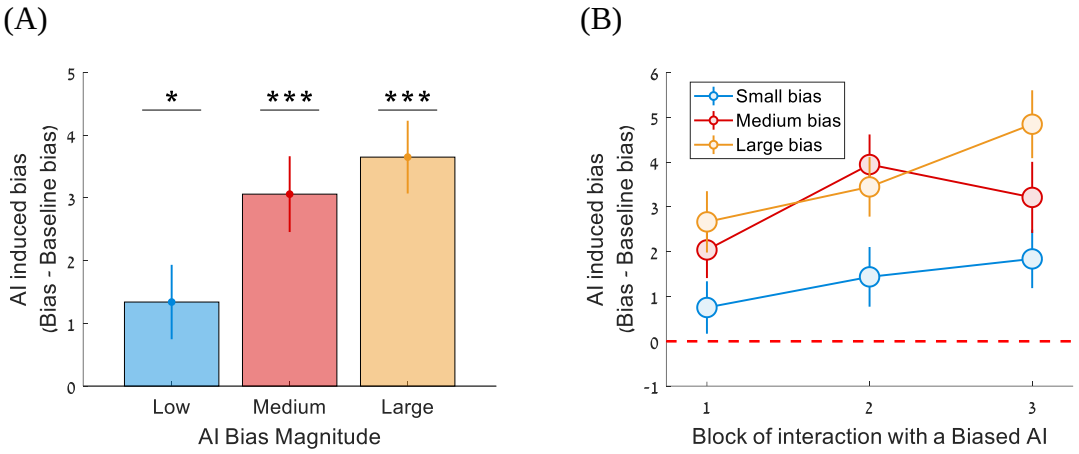


Fig. S2. AI induced bias as a function of the bias magnitude. (A) The AI induced bias was significantly higher than 0 for all levels of the AI magnitude bias (low, medium and high). (B) AI induced bias as a function of AI bias magnitude and block. The AI induced bias increased as a function of block.